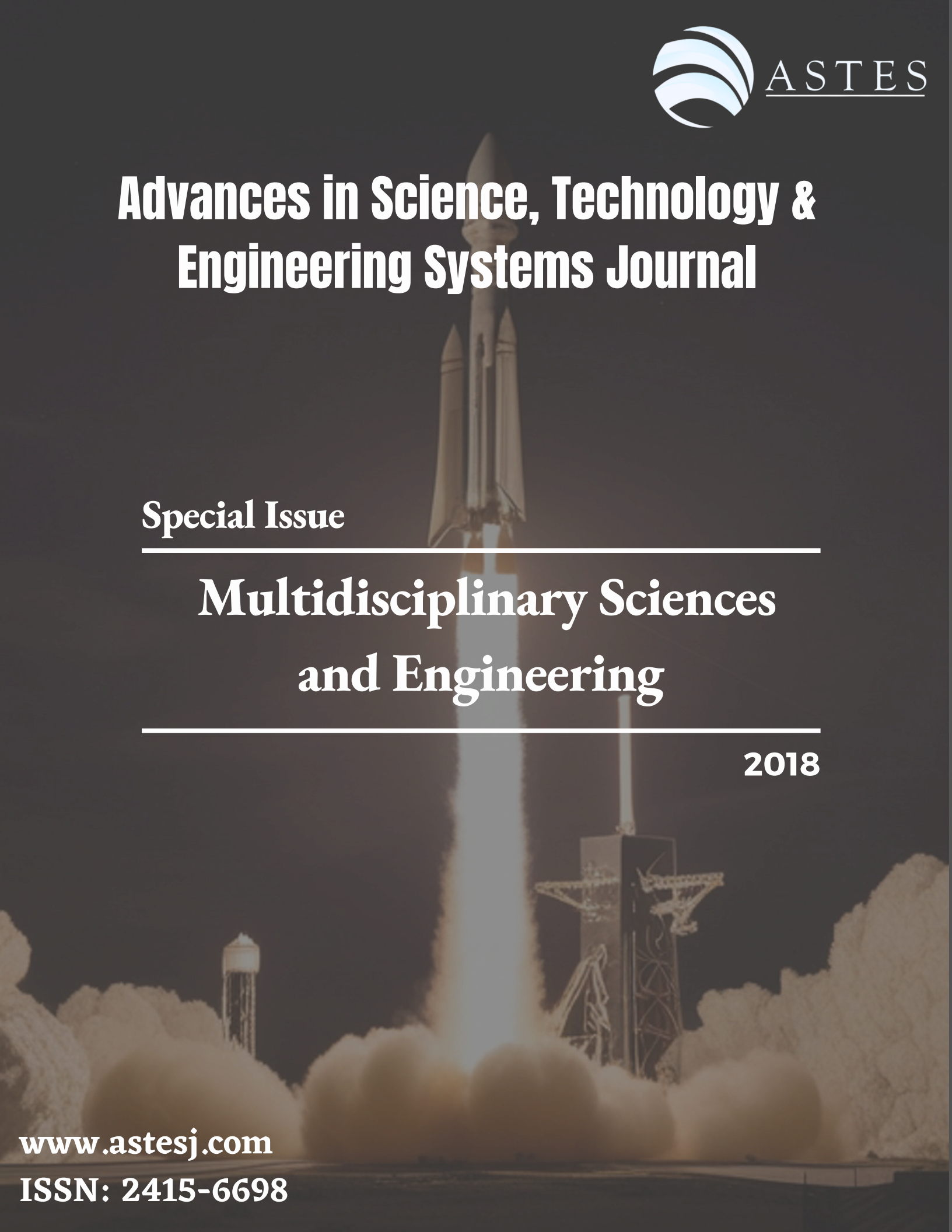


ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Multidisciplinary Sciences and Engineering

2018

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

Prof. Passerini Kazmerski
Pritzker School of Molecular Engineering, University of Chicago, USA

Guest Editors

Prof. Saleh Nabi
Mitsubishi Electric Research Laboratories, USA

Prof. Mohamed Abdelaziz Hassan Eleiwa
University of Hail, KSA

Dr. Mohammad Mohiuddin Uzzal
University of New Mexico, USA

Dr. Neeti Ananthaswamy
The Catholic University of America, USA

Dr. Arnab Hazari
University of Michigan, Ann Arbor, USA

Dr. Jules Ferdinand Pagna Disso
Nettitude Ltd, UK

Dr. Zamberi Bin Jamaludin
Universiti Teknikal Malaysia Melaka, Malaysia

Dr. Danielli A. Lima
Universidade Federal de Uberlândia, Brazil

Dr. Chenxiao Cai
Nanjing University of Science and Technology, China

Prof. María Jesús Espinosa Trujillo
Universidad Tecnológica Metropolitana, Mexico

Prof. Nicolae Tudoroiu
John Abbott College, Canada

Dr. Zupan Hu
University of Michigan, Ann Arbor, USA

Dr. Sarhan M. Musa
Prairie View A&M University, USA

Dr. Shahab Tayeb
University of Nevada, Las Vegas, USA

Dr. Bhuvan Modi
Prairie View A&M University, USA

Dr. Dillon Chrimes
Vancouver Island Health Authority, Canada

Dr. Sahar Mohamed Hamed
Canadian International College, Egypt

Dr. Shaddrack Yaw Nusenu
Koforidua Technical University, Ghana

Prof. Ahmad Yusairi Bani Hashim
Universiti Teknikal Malaysia Melaka, Malaysia

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

Prof. Jin Billy Li
University of Michigan, Ann Arbor, USA

Dr. Remya Ampadi Ramachandran
University of Illinois Chicago, USA

Dr. Mostafa Fadaee Fathabadi
Concordia University, Canada

Dr. Saurin H. Patel
Clemson University, USA

Dr. Matin Pirouz
University of Nevada, Las Vegas, USA

Dr. Raafat Elfouly
Rhode Island College, USA

Dr. Hossam Mosbah
Dalhousie University, Canada

Dr. Issa Tamer Elmabrouk Elfergani
Instituto de Telecomunicações, Portugal

Dr. Lihong Zheng
Charles Sturt University, Australia

Dr. Eman Hammad
University of Toronto, Canada

Dr. Nanthi Suthikarnnarunai
University of the Thai Chamber of Commerce, Thailand

Editorial

The Special Issue on *Multidisciplinary Sciences and Engineering (2018)* in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* presents a diverse and intellectually rich collection of research that reflects the growing importance of cross-disciplinary collaboration in addressing complex scientific and engineering challenges. The year 2018 witnessed an increasing convergence of knowledge domains, where traditional boundaries between sciences and engineering fields became less distinct, enabling more holistic and innovative approaches to problem-solving. This special issue captures that evolution by bringing together contributions that integrate concepts, methods, and tools from multiple disciplines to advance both theoretical understanding and practical applications.

A key theme emerging from this issue is the integration of scientific principles with engineering practices to develop comprehensive solutions. The included studies demonstrate how insights from physics, chemistry, mathematics, and computer science can be effectively combined with engineering methodologies to address real-world problems. This convergence has led to the development of more robust models, enhanced simulation techniques, and improved system designs. By leveraging multidisciplinary perspectives, researchers are better equipped to analyze complex systems, optimize performance, and innovate across a wide range of applications.

Another important aspect highlighted in this collection is the role of emerging technologies in facilitating multidisciplinary research. Advances in computational tools, data analytics, and digital platforms have significantly enhanced the ability of researchers to collaborate across disciplines. These technologies enable the integration of diverse datasets, the application of advanced analytical methods, and the development of intelligent systems that can adapt to dynamic conditions. Contributions in this issue illustrate how such technologies are being utilized in areas such as healthcare engineering, environmental systems, smart infrastructure, and advanced manufacturing.

The issue also emphasizes the importance of addressing global challenges through a multidisciplinary lens. Problems such as climate change, resource management, public health, and sustainable development require coordinated efforts that draw upon expertise from multiple fields. The research presented showcases innovative approaches that combine scientific inquiry with engineering solutions to create systems that are not only efficient and reliable but also socially and environmentally responsible. This holistic perspective is essential for developing solutions that are both impactful and sustainable in the long term.

Interdisciplinary education and research frameworks are also reflected in the works included in this issue. The contributions highlight the need for new educational models and research strategies that encourage collaboration, knowledge sharing, and integration across disciplines. By fostering a culture of multidisciplinary engagement, institutions and researchers can better prepare to tackle the increasingly complex challenges of the modern world. The studies presented demonstrate how such collaborative approaches lead to more innovative outcomes and a deeper understanding of interconnected systems.

The editorial team extends its sincere appreciation to all authors, reviewers, and contributors for their valuable efforts in making this special issue possible. Their commitment to excellence has ensured the publication of high-quality research that advances the frontiers of multidisciplinary sciences and engineering. The works presented in this issue not only reflect the current state of

research but also provide a strong foundation for future exploration, encouraging continued collaboration and innovation across diverse fields.

Guest Editor

Prof. María Jesús Espinosa Trujillo

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **March 2018**

CONTENTS

A Comparative Study of a Hybrid Ant Colony Algorithm MMACS for the Strongly Correlated Knapsack Problem
by Wiem Zouari, Ines Alaya and Moncef Tagina

Evaluation of Methods for Sentence Similarity for Use in Intelligent Tutoring System
by Emil Brajković, Daniel Vasić and Tomislav Volarić

Experimental Results and Numerical Simulation of the Target RCS using Gaussian Beam Summation Method
by Ghanmi Helmi, Khenchaf Ali, Pouliguen Philippe and Leye Papa Oussmane

The Chimera and the Cyborg
by Lev Lafayette, Bernd Wiebelt, Dirk von Suchdoletz, Helena Rasche, Michael Janczyk and Daniel Tosello

Harmonic Sensitivity Analysis of Power System with Wind Power
by Jignesh Patel and Satish Joshi

Simulation Strategy to Enhance Oxygen Uptake and Reaction Forces at Leg Joints and Vertebral Bodies During Ergometer Exercise Under Altered Gravity
by Yoshihiko Tagawa, Masayuki Omoto, Hiroo Matsuse and Naoto Shiba

Spectral Grid Impedance Identification on the Low-, Medium- and High-Voltage Level – System Design, Realization and Measurement Results of Grid Impedance Measurement Devices
by Hauke Wilken, Michael Jordan and Detlef Schulz

Design of True Random Numbers Generators with Ternary Physical Unclonable Functions
by Bertrand Francis Cambou

Improvement of Transmission Characteristics in Multilayer Dual Band Filter
by Ryosuke Nakano, Yuta Takeuchi, Mikio Tsuji and Hiroyuki Deguchi

Scientific Principles behind 2017 World Universiade of Safety Lantern and Torch
by Wei Long Chen and Fang Lin Chao

An Optical-based Fingertip Force Sensor
by Emanuele Lindo Secco, Taye F. Agidew and Atulya Kumar Nagar

Low-Dimensional Spaces for Relating Sensor Signals with Internal Data Structure in a Propulsion System
by Catherine Cheung, Nicolle Kilfoyle, Julio Valdés, Srishti Sehgal and Richard Salas Chavez

Stream Cipher by Reed-Solomon Codes

by Tao Wu and Ruomei Wang

Performance improvement of a wind energy system using fuzzy logic based pitch angle control
by Kanasottu Anil Naik, Chandra Prakash Gupta and Eugene Fernandez

Direct Torque Control of Multiphase Doubly Converter-fed Asynchronous Machines Incorporating the Harmonic Torques
by Javier Martínez-Román, Luis Serrano-Iribarnegaray, Juan Carlos del Pino-López and Pedro Cruz-Romero

Fuzzy Uncertainty Management in Multi-Shift Single-Vehicle Routing Problem
by Francesco Nucci

On The Development of a Reliable Gate Stack for Future Technology Nodes Based on III-V Materials
by Abhitosh Vais, Sonja Sioncke, Jacopo Franco, Vamsi Putcha, Laura Nyns, Arturo Sibaja-Hernandez, Rita Rooyackers, Sergio Calderon Ardila, Valentina Spampinato, Alexis Franquet, Jan Maes, Qi Xie, Michael Givens, Fu Tang, Xiang Jiang, Marc Heyns, Dimitri Linten, Jerome Mitard, Aaron Thean, Dan Mocuta and Nadine Collaert

Using Input Impedance to Calculate the Efficiency Numerically of Series-Parallel Magnetic Resonant Wireless Power Transfer Systems
by Thabat Thabet and John Woods

A Cyber-Vigilance System for Anti-Terrorist Drives Based on an Unmanned Aerial Vehicular Networking Signal Jammer for Specific Territorial Security
by Dhiman Chowdhury, Mrinmoy Sarkar and Mohammad Zakaria Haider

A detailed step-by-step electrical parameters identification method for photovoltaic generators using a combination of two approaches
by Selma Tchoketch Kebir, Mohamed Salah Ait Cheikh and Mourad Haddadi

Parametric Co-variance Assignment for a Class of Multivariable Stochastic Uncertain Systems: Output Feedback Stabilization Approach
by Qichun Zhang

Holistic Access Control and Privacy Infrastructure in Distributed Environment
by Uche Magnus Mbanaso and Gloria A Chukwudebe

Extending the Life of Legacy Robots: MDS-Ach via x-Ach
by Daniel M. Lofaro, Magdalena Bugajska and Donald Sofge

Using of the Flipped Classroom Learning with a Workshop Activity in Object-oriented Analysis and Design Course
by Pakawan Pugsee

Dc-link voltage drift compensation in a four level Double- Star Converter using redundant states via phase-shifted PWM strategy
by Carlos A. Reusser

An FPGA Implementation of Resource-Optimized Dynamic Digital Beamformer for a Portable Ultrasound Imaging System
by Jingwei Xu, Yi Zheng, Mohan Chirala and Mohamed Almekkawy

Effects of Dielectric Properties of the Material located inside Multimode Applicator on Microwave Efficiency
by Sofiya Ali Mekonnen, Sibel Yenikaya, Gökhan Yenikaya and Güneş Yilmaz

A Novel MICS Receiver with FSK Dual Band Demodulator
by Mouna Bettaieb, Saif Benali, Ghazi Bouzid and Hatem Trabelsi

Revealing Strengths, Weaknesses and Prospects of Intelligent Collaborative e-Learning Systems
by Amal Asselman, Azeddine Nasseh and Souhaib Aammou

Similarity-based Resource Selection for Scientific Workflows in Cloud Computing
by Takahiro Koita and Yu Manabe

A Machine Learning Framework Using Distinctive Feature Extraction for Hand Gesture Recognition
by Sudipta Saha, Aninda Saha, Zubayr Khalid, Pritam Paul and Shuvam Biswas

Interactive Virtual Reality Educational Application
by Shouq. Al Awadhi, Noor. Al Habib, Dalal Al-Murad, Fajer Al deei, Mariam Al Houti, Taha Beyrouthy and Samer Al-Kork

Visualizing Affordances of Everyday Objects Using Mobile Augmented Reality to Promote Safer and More Flexible Home Environments for Infants
by Miho Nishizaki

Towards Process Standardization for Requirements Analysis of Agent-Based Systems
by Khaled Slhoub and Marco Carvalho

It Takes Two to Tango: Merging Science and Creativity to Support Continued Innovation in the IoT Domain
by Helen Hasenfuss, Muftah Fraifer, Sameer Kharel, Asma Elmangoush, Alan Ryan and Walid Elgenaidi

Enhanced Ship Energy Efficiency by Using Marine Box Coolers
by Abdallah Aijjou, Lhoussain Bahatti and Abdelhadi Raihani

Theory Building of Quattro Bottom Line Approach for Sustainable Reverse Logistics from Government Perspective: The Indonesia Evidence
by Hesti Maheswari, Gatot Yudoko and Akbar Adhiutama

Evaluating the effect of Locking on Multitenancy Isolation for Components of Cloud-hosted Services
by Laud Charles Ochei and Christopher Ifeanyichukwu Ejiofor

A Novel Technique for Enhancing Color of Undersea Deblurred Imagery
by Chrispin Jiji and Nagaraj Ramrao

Virtual Output Queues Architecture for High Throughput Data Center Nodes
by Angelos Kyriakos, Ioannis Patronas, Georgios Tzimas, Vasileios Kitsakis and Dionysios Reisis

Extendable Board Game to Facilitate Learning in Supply Chain Management
by Linda William, Za'Aba Bin Abdul Rahim, Robert de Souza, Eko Nugroho and Rio Fredericco

An Electroencephalogram Analysis Method to Detect Preference Patterns Using Gray Association Degrees and Support Vector Machines
by Shin-ichi Ito, Momoyo Ito and Minoru Fukumi

Learning Personalization Based on Learning Style instruments
by Alzain Alzain, Steve Clark, Gren Ireson and Ali Jwaid

Designing Experiments: 3 Level Full Factorial Design and Variation of Processing Parameters Methods for Polymer Colors
by Jamal Al Sadi

An enhanced Biometric-based Face Recognition System using Genetic and CRO Algorithms
by Ola Surakhi, Mohammad Khanafseh and Yasser Jaffal

FPGA Implementation of Ultra-High Speed and Configurable Architecture of Direct/Inverse Discrete Wavelet Packet Transform Using Shared Parallel FIR Filters
by Mouhamad Chehaitly, Mohamed Tabaa, Fabrice Monteiro, Juliana Srour and Abbas Dandache

Developing A Conceptual Framework of Product-Service System Management Toward Firms' Sustainability for Indonesian Industrial Estate Firms
by Christina Wirawan, Gatot Yudoko and Yuliani Dwi Lestari

A Method for Generating, Evaluating and Comparing Various System-level Synthesis Results in Designing Multiprocessor Architectures
by Peter Arato and Gyorgy Racz

Acoustic Signal Processing and Noise Characterization Theory via Energy Conversion in a PV Solar Wall Device with Ventilation through a Room
by Himanshu Dehra

SDN-based Network Control Method for Distributed Storage Systems
by Luis Guillen, Satoru Izumi, Toru Abe, Hiroaki Muraoka and Takuo Suganuma

A New Study Performance Control of PMSMs: Validity Abacus Approach
by Sabrine Jebri and Khaled Nouri

Automating Hostel Telephone Systems

by Rohan Prabhu Murje, Bhaskar Rishab, Krishna Gopalrao Jorapur, MuccatiraThimmaiah Karumbaiah and Muddenahalli Nagendrappa Thippeswamy

Evolving AL-FEC Application on 5G NGMN-Edge Computing Systems
by Christos Bouras and Nikolaos Kanakis

Fuzzy Logic Based Selective Harmonic Elimination for Single Phase Inverters
by Zeynep Bala Duranay and Hanifi Guldemir

Implementation of a Robot Contest for Distance Education
by Eiken Lübbers and Dierk Schoen

Influence of Torrefaction on Gasification of Torrefied Palm Kernel Shell
by Razi Ahmad, Mohd Azlan Mohd Ishak, Khudzir Ismail and Nur Nasulhah Kasim

Complexity Drivers in Digitalized Work Systems: Implications for Cooperative Forms of Work
by Benedikt Andrew Latos, Markus Harlacher, Florens Burgert, Verena Nitsch, Philipp Przybysz and Susanne Mütze-Niewöhner

Supporting Active Aging Through A Home Automation Infrastructure for Social Internet of Things
by Vittorio Miori, Dario Russo and Luca Ferrucci

Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis
by Charles Frank, Asmail Habach, Raed Seetan and Abdullah Wahbeh

NemoMap: Improved Motif-centric Network Motif Discovery Algorithm
by Tien Huynh, Somadina Mbadiwe and Wooyoung Kim

Design of smart chess board that can predict the next position based on FPGA
by Alaa Hamza Omran and Yaser Muhammad Abid

Guidance Law Based on Line-of-Sight Rate Information Considering Uncertain Modeled Dynamics
by Saori Nakagawa, Takeshi Yamasaki, Hiroyuki Takano and Isao Yamaguchi

Attitude Control Simulation of a Legged Aerial Vehicle Using the Leg Motions
by Yoshiyuki Higashi and Soonki Chang
An Experimental Investigation of Reconfigurable UWB Modified Octagonal Microstrip Monopole Patch Antenna with Switchable and Tunable Band-Notched Characteristic
by Ruchi Paliwal, Ruchi Paliwal, Rajesh Kumar Singh and Shiban Kishen Koul

Psychoacoustical Approach in Soundscape Characterization
by Mia Suhanek, Ivan Djurek, Sanja Grubeša and Antonio Petošić

Analysis of Drain Current Transient Response of Gate Pulse Voltage in AlGaN / GaN High Electron Mobility Transistors
by Hirohisa Taguchi, Kazuto Akahori, Takuma Shimazu and Honoka Tanabe

Development and Testing of Intelligent Wheelchair Controller for Quadriplegic Patients
by Mohammed Faeik Ruzaij Al-Okby, Sebastian Neubert, Norbert Stoll and Kerstin Thurow

A Dual-Band 90-Degree SiGe HBT Active Phase Shifter Based on Band-Pass and Band-Stop Designs Using Dual-Band Resonators
by Yasushi Itoh

Contactless Power and Bidirectional Data Transmission via Magnetic Field
by Jia-Jing Kao, Chun-Liang Lin, Chih-Cheng Huang and Hau-Shian Jian

Computational Techniques to Recover Missing Gene Expression Data
by Negin Fraidouni and Gergely Zaruba

Towards an Efficient Federated Cloud Service Selection to Support Workflow Big Data Requirements
by Mohamed Adel Serhani, Hadeel Al Kassabi and Ikbal Taleb

A novel mixed-mode universal biquad employing plus current output DVCCs
by Takao Tsukutani, Yasutomo Kinugasa and Noboru Yabuki

An Overview of Data Center Metrics and a Novel Approach for a New Family of Metrics
by Moises Levy and Daniel Raviv

Pole-Changing Windings for Close Ratio and 1:N Ratio Using the 3//Y / 3//Y Method
by Gabor Kovacs

Closed Approach of a Decoder Mobile for the 406 Mhz Distress Beacon
by Billel Ali Srihen, Jean-Paul Yonnet and Malek Benslama

A Ubiquitous and Configurable Wrist-Worn Sensor Node in Hazardous Gases Detection
by Mostafa Haghi, Kerstin Thurow and Regina Stoll

Frameworks for Performing on Cloud Automated Software Testing Using Swarm Intelligence Algorithm: Brief Survey
by Mohammad Hossain, Sameer Abufardeh and Sumeet Kumar

Wireless Channel Measurement and Modeling in Industrial Environments
by Kun Zhang, Liu Liu, Cheng Tao, Ke Zhang, Ze Yuan and Jianhua Zhang

What Should Be Considered for Acceptance Mobile Payment: An Investigation of the Factors Affecting of the Intention to Use System Services T-Cash
by Riyan Rizkyandy, Djoko Budiyanto Setyohadi and Suyoto

Developing Students' Motivation for Learning through Practical Problems in School
by Mikhail Rodionov and Zhanna Dedovets

Student Performance Evaluation Using Data Mining Techniques for Engineering Education
by Veena Deshmukh, Srinivas Mangalwede and Dandina Hulikunta Rao

Topology, Discontinuities and Dimension Effects on CMOS Rotary Traveling Wave Oscillators
by Rogelio Manuel Higuera Gonzalez, Mónico Linares Aranda, Luis Hernández Martínez and Carlos Sanabria Díaz

Parallelizing Combinatorial Optimization Heuristics with GPUs
by Mohammad Harun Rashid and Lixin Tao

A Study on Improving Security and Efficiency using Dynamic Ownership Management in Client-Side Deduplication Environments
by Won-Bin Kim and Im-Yeong Lee

Sequential Card Activities of Recollect Experience for Elderly with Dementia
by Fang Lin Chao and Boxiu Fanjiang

Steganography System with Application to Crypto-Currency Cold Storage and Secure Transfer
by Michael J. Pelosi, Nimesh Poudel, Pratap Lamichhane and Danyal Badar Soomro

The Impacts of Distribution Generators' Size and Location on Power Efficiency and Voltage Profile in Radial LV Networks
by Maher M.A Al-Maghalseh

Probabilistic Method for Anomalies Detection Based on the Analysis of Cyber Parameters in a Group of Mobile Robots
by Elena Basan, Alexander Basan and Oleg Makarevich

Design and Implementation of Closed-loop PI Control Strategies in Real-time MATLAB Simulation Environment for Nonlinear and Linear ARMAX Models of HVAC Centrifugal Chiller Control Systems
by Nicolae Tudoroiu, Mohammed Zaheeruddin, Songchun Li and Elena-Roxana Tudoroiu

The experience of implementation with Agile Business Process Management
by Denis Rodríguez and Enrique Silva Molina

An Empirical Study of Icon Recognition in a Virtual Gallery Interface
by Denise Ashe, Alan Eardley and Bobbie Fletcher

The Model Development of an Effective Triggering System of Production Kanban Size towards Just-In-Time (JIT) Production
by Mohd Norzaimi Che Ani, Shahrul Kamaruddin and Ishak Abdul Azid

A Model of EM Fields from Static Data for Moving Conducting Cylinder
by Esmail Mohamed Mohamed Abuhdima and Robert Prewitt Penno

Synthesis of QDI Combinational Circuits using Null Convention Logic Based on Basic Gates
by Duarte Lopes de Oliveira, Orlando Verducci, Vitor Leandro Vieira Torres, Robson Moreno and Lester de Abreu Faria

Which User of technology? Perspectivising the UTAUT model by application of the SFL language Pronoun System towards a systems perspective of technology acceptance and use

by Cheryl Marie Cordeiro

Protecting Private Data using Improved Honey Encryption and Honeywords Generation Algorithm
by Thanda Win and Khin Su Myat Moe

A Comparative Analysis of two Controllers for Trajectory Tracking Control: Application to a Biological Process
by Abyad Mohamed, Karama Asma and Khallouq Abdelmounaim

Metaheuristics for Solving Facility Location Optimization Problem in Lagos, Nigeria
by Chika Yinka-Banjo and Babatunde Opesemowo

The Value of Integrating MSRP Protocol in E-learning Platforms of Universities
by Khalifa Sylla, Samuel Ouya, Masamba Seck and Gervais Mendy

Minimizing Time Delay of Information Routed Across Dynamic Temporal Sensor Networks
by Michael J. Hirsch and Azar Sadeghnejad

Q-Learning versus SVM Study for Green Context-Aware Multimodal ITS Stations
by Adel Mounir Said, Emad Abd-Elrahman and Hossam Afifi

Efficient Alignment of Very Long Sequences
by Chunchun Zhao and Sartaj Sahni

Machine Learning Applied to GRBAS Voice Quality Assessment
by Zheng Xie, Chaitanya Gadepalli, Farideh Jalalinajafabadi, Barry M.G. Cheetham and Jarrod J. Homer

MRI images Enhancement and Brain Tumor Segmentation
by Aye Min and Zin Mar Kyu

Semi-Autonomous Robot Control System with an improved 3D Vision Scheme for Search and Rescue Missions. A joint research collaboration between South Africa and Argentina

MARWIN: Localization of an Inspection Robot in a Radiation exposed Environment
by Andre Dehne, Nantwin Moller and Thorsten Hermes

Modified HOG Descriptor-Based Banknote Recognition System
by Tamarafinide Victory Dittimi and Ching Yee Suen

Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images
by Margaret Lech, Melissa Stolar, Robert Bolia and Michael Skinner

Characteristic, Thermochemical Behaviors and Kinetic of Demineralized and Torrefied Empty Fruit Bunches (EFB)
by Nur Nasulhah Kasim, Khudzir Ismail, Alina Rahayu Mohamed, Mohd Azlan Mohd Ishak, Razi Ahmad and Wan Izhan Nawawi Wan Ismail

IR Sensing Embedded System Development for Prototype Mobile Platform and Multisensory Data Fusion for Autonomous Convoy
by Hubert Bryan Riley, David Solomon Raj Kondru and Mehmet Celenk

Virtual Watershed System: A Web-Service-Based Software Package For Environmental Modeling
by Rui Wu, Connor Scully-Allison, Moinul Hossain Rifat, Jose Painumkal, Sergiu Dascalu and Jr. Frederick C Harris

Robot Self-Detection System
by Ivaylo Penev, Milena Karova, Mariana Todorova and Danislav Zhelyazkov

Elliptical Printed Dipole Antenna Design using ANN Based on Levenberg–Marquardt Algorithm
by Ali I. Hammoodi, Mariofanna Milanova and Haider Raad

The Effect of Fluence Variations of Nd:YAG Laser Ablation and Sample Condition on Human Tooth
by Fatanah Mohamad Suhaimi, Nur Zarifha Zainol Alam, Suriani Mat Ariffin, Nurul Atiqah Abd. Razak and Mohammad Khairul Azhar Abdul Razab

An Approach for Determining Rules used to Select Viable Junction Design Alternatives Based on Multiple Objectives
by Erwin Bezembinder, Luc Wismans and Eric van Berkum

Improved DTC Control Strategy of B12 Inverter Fed BLDC Motor Drives Considering Commutation Torque Dips
by Rabiaa Mars, Badii Bouzidi, Bassem El Badsi and Abderrazak Yangui
An Integrated & Secure System for Wearable Devices
by Callum Owen-Bridge, Stewart Blakeway and Emanuele Lindo Secco

Using Nullors to Modify Linear Model Parameters of Transistors in an Analog Circuit
by Reza Hashemian

Economic and Social Sustainability for Iraqi Middle Provinces
by Maysoon Abdullah Mansor

Minimally Invasive, Thermal Energy Based, Cost-Efficient Method to Measure Fluid Flows in Compact Systems
by Saurin Patel and Rick Walker

Simulation-Optimisation of a Granularity Controlled Consumer Supply Network Using Genetic Algorithms
by Zeinab Hajiabolhasani, Romeo Marian and John Boland

Numerical Simulation of One Pavement Structure of Polyethylene Terephthalate Submitted to Static Point Loads
by Juan Sebastián Gilart Pulido, María Camila Zambrano Bello and Ender Jhobany Orduz Duarte

Multi-Objective Path Optimization of a Satellite for Multiple Active Space Debris Removal Based on a Method for the Travelling Serviceman Problem
by Masahiro Kanazaki, Yusuke Yamada and Masaki Nakamiya

Perfect Molding Challenges and The Limitations “A Case Study”
by Tan Lay Tatt, Lim Boon Huat, Rosli Muhammad Tarmizi and T. Joseph Sahaya Anand

Towards Adoption of Authentication and Authorization in Identity Management and Single Sign On
by Ujjwal Joshi, Sangwhan Cha and Saeed Esmaili-Sardari

Emotional Impact of Suicide on Active Witnesses: Predicting with Machine Learning
by Richard Osei Agjei, Emmanuel Awuni Kolog, Daniel Dei and Juliet Yayra Tengey

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, XX-YY (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Using Input Impedance to Calculate the Efficiency Numerically of Series-Parallel Magnetic Resonant Wireless Power Transfer Systems

Thabat Thabet[*1,2], John Woods[1]

[1]*Essex University, School of Computer Science and Electronic Engineering, CO4 3SQ, UK*

[2]*Technical College of Mosul, Computer Engineering Department, Mosul, Iraq*

ARTICLE INFO

Article history:
Received : 20 March, 2018
Accepted : 27 April, 2018
Online :7 May, 2018

Keywords :
wireless power transfer
magnetic resonant
maximum efficiency
series-parallel connection
SP
input impedance

ABSTRACT

Analysis of magnetic resonant wireless power transfer systems aims to achieve maximum efficiency of the power transfer. From the analysis we wish to derive the maximum power and the frequency at which this occurs. This paper presents a method to estimate these two required values and to achieve this requires the solution of the input impedance equation numerically. The frequency of the maximum efficiency is found when the imaginary of the input impedance is close to zero, and it could be different to the natural resonant frequency. We estimate the efficiency value which depends on the real value part of the input impedance. The proposed method has been applied to one of the four types of possible connections; a series-parallel (SP) connection although similar approaches could be applied to the others. In some cases the maximum efficiency shifts away from the resonant frequency. Therefore, this paper shows how to use the same equations to achieve maximum efficiency at resonance and suggests a design method to achieve this practically.

1. Introduction

This paper is an extension of work originally presented in 2017 IEEE International Conference on Circuits, System and Simulation [1]. Where the impact of connection type on the efficiency of magnetic resonant wireless power transfer systems is studied. There are four potential types of connection in the wireless system due to the resonance type in the transmitter and the receiver; they are: series-series (SS); series-parallel (SP); parallel-series (PS); and parallel-parallel (PP). The equivalent circuit of each type can be expressed by applying Kirchhoff's law of voltage on each loop in the circuit producing a set of equations [1]. This is called the impedance matrix representation. The effect of changing the load resistor and the gap between the two coils is studied for each connection.

As a near field technology, magnetic resonant wireless power transfer systems can achieve high efficiency in specific conditions. Where the efficiency of the maximum transfer of power is affected by several factors; these are the resonant frequency, the parameters of the coils (inductance, size, and shape), the distance between them and so the mutual inductance, the load resistor, and the connection type of the transmitter and the receiver [1],[2]. To improve the performance of the wireless system, it is important to calculate the efficiency and understand how it is affected by each factor.

Generally, there are two main methods to analyse the system. These methods are coupled mode theory (CMT) and circuit theory (CT) [3]. In the former method (CMT), the system can be described as differential equations [4], [5], [6]. However, according to [7] this method is complicated, undesirable and inconvenient. Therefore researchers tend to use the equivalent circuit for magnetic coupled systems.

The second method (CT) is more familiar to circuit designers than the coupled-mode theory used by the physicists. In this method, the magnetic coupling is represented as the mutual inductance between the two coils. The equivalent circuit of the wireless power transfer system can be formulated as a two port network, which can be represented by either impedance matrix or

[*]Corresponding Author, tfytha@essex.ac.uk

scattering matrix [5]. Under this main concept of (CT), many researchers have used different modes with different names to study magnetic resonant wireless power transfer systems, such as magnetic resonant coupling (MRC) [7], [8], [9], and reflected load theory (RLT) [4], [10], [11], [12].

The (MRC) method is used to study magnetic resonant wireless power transfer systems by applying a scattering matrix [7], [8], [13]. They use this method because the practical measurement of scattering parameters is more convenient than other methods at high frequencies.

The (RLT) method is commonly used to analyse transformers by electrical engineers, and it can be used to analyse magnetic resonant wireless power transfer systems. The method states that the current in the transmitter coil is dependent on the load in the receiver coil, and the reflected load in the transmitter is not always the same as the actual load [4].

Starting with the impedance matrix representation to express the wireless system, this paper presents another mode to calculate the maximum efficiency of the transfer power and its associated frequency which can be different than the resonant frequency in some cases. Series-parallel type connection has been chosen to examine the method and calculate the required values. The paper also includes an explanation of the previous mode developed in [1]. The results of the two modes are compared in order to evaluate the new mode and show the differences in the calculation methods.

2. Theoretical Analysis

The transmitter of series-parallel magnetic resonant wireless power transfer system consists of a coil (L_1) in serial with a capacitor (C_1). While the receiver coil (L_2) is in parallel with the capacitor (C_2), as shown in Fig. 1. The two capacitors are chosen to make the two coils resonate at the same frequency. The voltage source (V_S) of the system has an internal resistance of (R_S); and R_L is the load resistor. The circuit can be presented as follows:

$$\begin{bmatrix} V_S \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} R_S + X_{L1} + X_{C1} & -X_M & 0 \\ -X_M & X_{L2} + X_{C2} & -X_{C2} \\ 0 & -X_{C2} & R_L + X_{C2} \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} \quad (1)$$

Where $XL=2\pi fL$ and $XC=1/2\pi fC$; f is the frequency and f_o is the selected frequency for resonance. Equation (1) can be rewritten in a different way.

$$[I] = [inv\ Z][V] \quad (2)$$

Through the following steps, the efficiency of the maximum transfer power can be found:

$$P_{out} = I_3^2\ R_L = (inv\ Z(3,1)\ V_S)^2 R_L \quad (3)$$

$$P_{in} = \left(\frac{V_S}{R_S+R_L}\right)^2 R_L \quad (4)$$

$$eff = inv\ Z(3,1)^2\ (R_s + R_L)^2 * 100\% \quad (5)$$

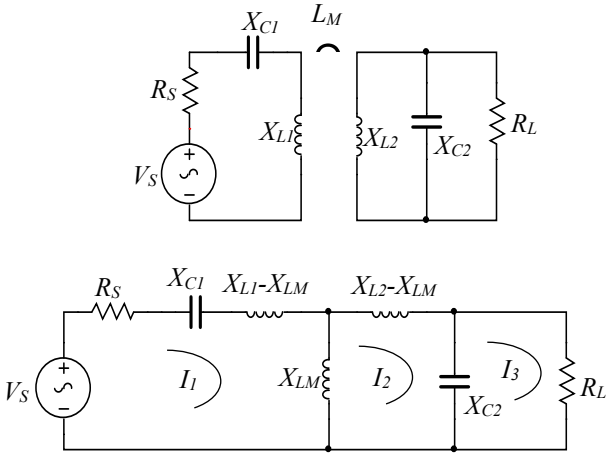


Figure 1: Equivalent circuit of magnetic resonant coupling series-parallel antennas in two forms

3. New Mode of Efficiency Calculation

Another way to calculate the efficiency is as follows:

$$V_S = I_1\ (R_S + Z_{in}) \quad (6)$$

It can be found from equation (2) that:

$$inv\ Z(1,1) = \frac{I_1}{V_S} \quad (7)$$

$$Z_{in} = \frac{1}{inv\ Z(1,1)} - R_S \quad (8)$$

$$Z_{in} = R_{in} + jX_{in} \quad (9)$$

Where R_{in} and X_{in} are the real and the imaginary parts of the input impedance, respectively.

Firstly, it is observed that maximum efficiency occurs when the imaginary part of the input impedance is equal to zero. Therefore, the frequency of maximum efficiency (even if it is different than the original resonant frequency) can be found from:

$$X_{in}(f) = 0 \quad (10)$$

$$X_{in}(f) = Af + \frac{B}{f} + Cf^3 + \frac{D}{f^3} \quad (11)$$

The values of A, B, C, and D depend on the parameters of the circuit. It is apparent that the equation for X_{in} is complicated and does not have an analytic solution. Therefore, it has to be solved numerically in order to calculate the required frequency. The

second step is to find the value of the maximum efficiency which is achieved by finding the real part R_{in} at that calculated frequency. The equivalent circuit in this stage is shown in Fig.2. Assuming maximum transfer of power the maximum efficiency (*eff*) is derived as:

$$eff = \frac{4\,R_S\,R_{in}}{(R_S+R_{in})^2} * 100\% \qquad (12)$$

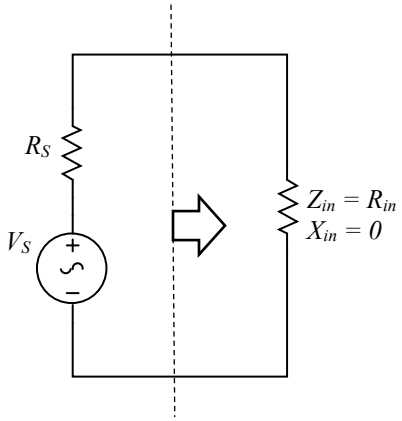


Figure 2: Equivalent circuit at the maximum transfer of power

4. Results and Discussion

The representation of the series-parallel circuit is modelled in Matlab to calculate the mutual inductance and the efficiency of the two modes. To support the theoretical results, several experiments have been conducted. In this work a set of two circular pancake coils are used in all experiments as the primary and secondary. Each coil has 2.9 cm inner radius and 8 cm outer, with 17 turns and an inductance equal to 31.4µH. In addition to that, a set of capacitors equal to 185pF, tuned to work at 2.088MHz were connected to the coils in series in the transmitter and in parallel in the receiver.

4.1. Effect of Distance

Starting with a 50Ω load resistor, the efficiency is found at different gaps between the two coils 7 cm, 5 cm, and 2 cm. Figure 3 shows efficiency versus frequency calculated by equation (5). It is clear from the figure that the maximum efficiency increases and shifts away from the original resonant frequency for small gaps. The table contains the maximum efficiency and the associated frequency calculated by equations (10-12) for the same gaps between the two coils.

Table: Maximum efficiency and associated frequency calculated by the input impedance for the different cases

Gap (cm)	2	5	7	5	5
R_L (Ω)	50	50	50	100	1k
$f\,(X_{in}\approx 0)$ (MHz)	2.43	2.14	2.11	2.14	2.17
$R_{in}\,(X_{in}\approx 0)$ (Ω)	13	2.35	0.87	4.7	45.77
eff_{max} (%)	65.5	17.15	6.7	31.4	99.8

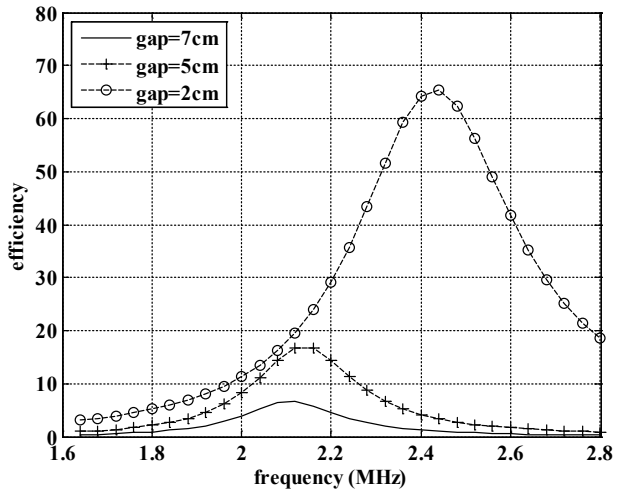


Figure 3: Efficiency versus frequency at different gaps,

4.2. Effect of Load Resistor

Working on different load resistor values affects the efficiency of the wireless system. Figure 4 shows the system has high efficiency with small load at the same distance. In all cases there is a small shift in the maximum efficiency point compared to the resonant frequency.

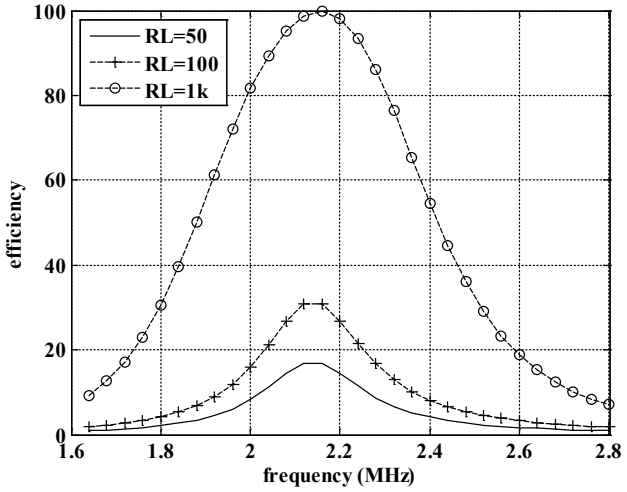


Figure 4: Efficiency versus frequency at different R_L, gap=5cm

In order to show the shift in maximum efficiency, the efficiency and input impedance curves including the real and the imaginary parts for the last case in the Table, all are shown in Fig. 5. The figure also defines the original resonant frequency at 2.088MHz.

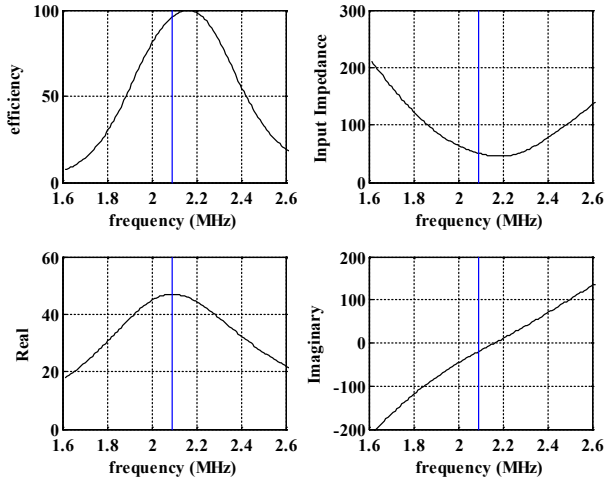


Figure 5: Efficiency, input impedance, real, and imaginary versus frequency, R_L= 1kΩ, gap=5cm

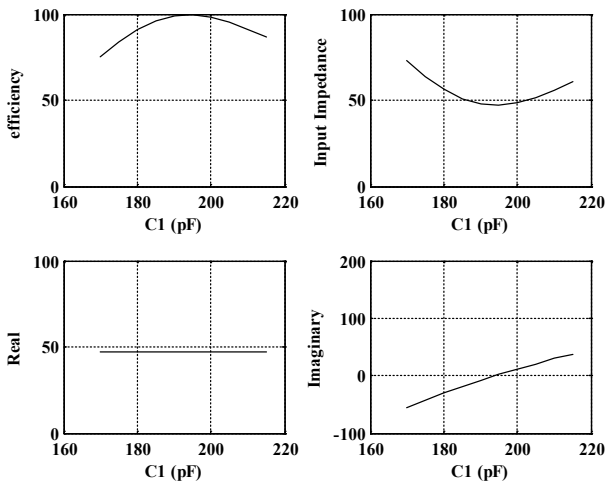


Figure 6: Efficiency, input impedance, real, and imaginary versus C_1, at f_0=2.088MHz, R_L= 1kΩ, gap=5cm

The maximum efficiency and the associated frequency (calculated by the input impedance and included in the Table) are well matched; the curves are shown in Fig. 3-5.

5. Moving the Maximum Efficiency to the Resonant Frequency

For feasible and efficient systems, it is important to obtain maximum efficiency at the resonant frequency. Otherwise, the system might lose a significant part from the transfer power. It is obvious from the previous figures that there is no symmetry in the curves, and the maximum efficiency shifts to one side of the resonant frequency. It can be concluded that tuning one of the capacitors is sufficient to re-merge the maximum efficiency with the original resonant frequency. This is possible by using the input impedance equation (the imaginary part) in a different way:

$$X_{in}(C1) = 0 \tag{13}$$

Using the resonant frequency to solve the last equation provides the capacitor value C_1 that achieves the maximum efficiency at that frequency, as shown in Fig. 6. The figure shows that C_1 =195pF is the required value; it can be calculated numerically from equation (13). Using the calculated value of C_1 equates the imaginary part of the input impedance to zero and then matches the highest efficiency with the resonant frequency, as shown in Fig.7.

Applying the proposed method empirically to improve the performance of the system by moving the maximum efficiency to the original resonant frequency is achieved by measuring the phase shift between the real and the imaginary parts of the input impedance. The phase shift indicates the imaginary part; its zero value means that the imaginary part of the input impedance is zero and this is the required value from the design, according to the following formula:

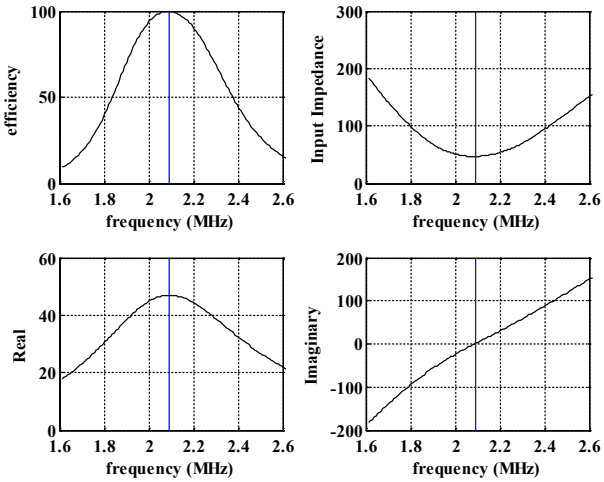


Figure 7: Efficiency, input impedance, real, and imaginary versus frequency, R_L= 1kΩ, gap=5cm, C_1=195pF

$$phase\ shift = tan^{-1}\frac{X_c}{R} \tag{14}$$

The suggested design is shown in Fig.8. It is suitable for specific ranges in the series-parallel parameters, where a capacitive effect is observed in the input impedance of the wireless system. To reduce this effect, increasing of the capacitor in the transmitter side is needed. The figure shows that the capacitor in the transmitter side is tuned according to the calculation of the phase shift of the input impedance.

To design a system theoretically the steps are: specify the parameters of the circuit; choose the resonant frequency; and calculate the required value of the tuning capacitor to achieve maximum efficiency at resonance. To apply the system practically requires the following steps: calculate the phase shift of the input impedance; automatically tune the transmitter capacitor to reduce the phase shift; and then reduce the imaginary part of the input impedance.

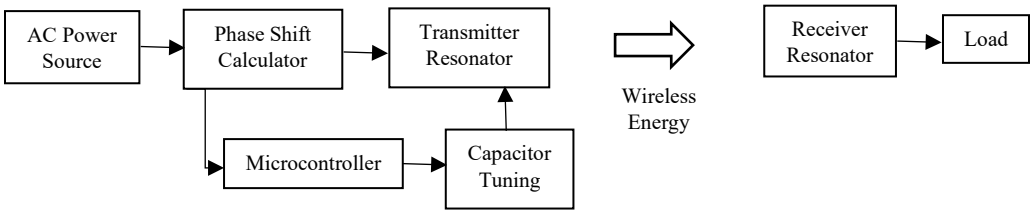


Figure 8: Block diagram of the suggested design

6. Conclusion

Having more than one mode to calculate efficiency is more convenient and allows results to be compared to one another.

The proposed method is a quick and simple mode to identify i) the maximum efficiency in a series-parallel wireless power transfer system and ii) the associated frequency. The idea is to equate the imaginary part of the input impedance with zero to find the frequency; and then calculate the maximum efficiency at that frequency from the real part.

Initially, the proposed method was applied to the series-parallel case. In our future work, this method can be applied on the other topologies of resonant wireless power systems.

This work markedly improves the efficiency of a series-parallel wireless system. The limitations of achieving this practically is limited to specific ranges of parameters where a capacitive effect is observed in the input impedance.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] T. Thabet, and J. Woods, "Impact of connection type on the efficiency of wireless power transfer systems.", in 7th IEEE International Conference on Circuits, System and Simulation (ICCSS), 2017. https://doi.org/10.1109/CIRSYSSIM.2017.8023200

[2] T. Thabet, and J. Woods, "An Approach to Calculate the Efficiency for an N-Receiver Wireless Power Transfer System," IJACSA, 6(9), 91-98, 2015. https://doi.org/10.14569/IJACSA.2015.060912

[3] X. Wei, Z. Wang, and H. Dai, "A critical review of wireless power transfer via strongly coupled magnetic resonances," Energies, 7(7), 4316-4341, 2014. https://doi.org/10.3390/en7074316

[4] O. Rönnbäck, "Optimization of Wireless Power," Master Thesis, Luleå University of Technology, 2013.

[5] A. Bodrov, and S.-K. Sul, "Analysis of wireless power transfer by coupled mode theory (CMT) and practical considerations to increase power transfer efficiency," InTech, 2012. https://doi.org/10.5772/25270

[6] H. Shim, J. Park, S. Nam, and B. Lee, "A criterion Proposed for Inductive Coupling and Magnetic Resonance Coupling in Wireless Power Transfer System." in 26th Asia-Pacific Microwave Conference (APMC), 2014. http://ieeexplore.ieee.org/document/7067678/

[7] T. Beh, M. Kato, T. Imura, and Y. Hori, "Wireless Power Transfer System via Magnetic Resonant Coupling at Fixed Resonance Frequency—Power Transfer System Based on Impedance Matching—," EVS-25 Shenzhen, China, 2010. https://pdfs.semanticscholar.org/7c5c/1dca95a922971ce5a7e28700876ff29b6448.pdf

[8] T. Imura, H. Okabe, and Y. Hori, "Basic experimental study on helical antennas of wireless power transfer for electric vehicles by using magnetic resonant couplings." in 5th IEEE Vehicle Power and Propulsion Conference, 2009. https://doi.org/10.1109/VPPC.2009.5289747

[9] T. C. Beh, M. Kato, T. Imura, S. Oh, and Y. Hori, "Automated impedance matching system for robust wireless power transfer via magnetic resonance coupling," IEEE Transactions on Industrial Electronics, 60(9), 3689-3698, 2013. https://doi.org/10.1109/TIE.2012.2206337

[10] M. Kiani, U.-M. Jow, and M. Ghovanloo, "Design and optimization of a 3-coil inductive link for efficient wireless power transmission," IEEE transactions on biomedical circuits and systems, 5(6), 579-591, 2011. https://doi.org/10.1109/TBCAS.2011.2158431

[11] M. Kiani, and M. Ghovanloo, "The circuit theory behind coupled-mode magnetic resonance-based wireless power transmission," IEEE Transactions on Circuits and Systems I, 59(9), 2065-2074, 2012. https://doi.org/10.1109/TCSI.2011.2180446

[12] K. S. Kiran, S. Brahma, S. Parida, and R. Behera, "Analysis of inductive resonant coupled WPT system using Reflected Load Theory." in IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), 2014. https://doi.org/10.1109/PEDES.2014.7041987

[13] T. Imura, and Y. Hori, "Maximizing air gap and efficiency of magnetic resonant coupling for wireless power transfer using equivalent circuit and neumann formula," IEEE Transactions on Industrial Electronics, 58(10), 4746-4752, 2011. https://doi.org/10.1109/TIE.2011.2112317

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 01-22 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

A Comparative Study of a Hybrid Ant Colony Algorithm MMACS for the Strongly Correlated Knapsack Problem

Wiem Zouari*, Ines Alaya, Moncef Tagina

COSMOS Laboratory, National School of Computer Science University of Manouba, 2010, Tunisia

ARTICLEINFO

Article history:
Received: 15 August, 2018
Accepted: 22 October, 2018
Online: 01 November, 2018

Keywords:
Ant colony optimization
Hybrid metaheuristic
Stochastic greedy approach
Local search
Strongly Correlated Knapsack Problem
Empirical study

ABSTRACT

Metaheuristic hybridization has recently been widely studied and discussed in many research works as it allows benefiting from the strengths of metaheuristics by combining adequately the different algorithms. MMACS is a new hybrid ant colony optimization algorithm based on the foraging behavior of ants. This algorithm presents two hybridization levels. The first hybridization consists in integrating the Ant Colony System selection rule in MAX-MIN Ant System. The second level of hybridization is to combine the hybridized ACO and an algorithm based on a local search heuristic, then both algorithms are operating sequentially. The optimal performance of MMACS algorithm depends mainly on the identification of suitable values for the parameters. Therefore, a comparative study of the solution quality and the execution time for MMACS algorithm is presented. The aim of this study is to provide insights towards a better understanding of the behavior of the MMACS algorithm with various parameter settings and to develop parametric guidelines for the application of MMACS to the Strongly Correlated Knapsack Problems. Results are compared with well-known Ant Colony Algorithms and recent methods in the literature.

1 Introduction

MMACS algorithm is a hybrid metaheuristic that was proposed in a previous work [1] and employed to solve one of the most complex variants of the knapsack problem which is the Strongly Correlated Knapsack Problem (SCKP). The proposed approach combines a proposed Ant Colony Optimization algorithm (ACO) with a 2-opt algorithm. The proposed ACO scheme combines two ant algorithms: the MAX-MIN Ant System and the Ant Colony System. At a first stage, the proposed ACO aims to solve the SCKP to optimality. In case an optimal solution is not found, a proposed 2-opt algorithm is used. Even if the 2-opt heuristic fails to find the optimal solution, it would hopefully improve the solution quality by reducing the gap between the found solution and the optimum.

An optimal resolution of a combinatorial optimization problem by applying an approximate method requires an adequate balance between exploitation of the best available solutions and wide exploration of the research space. On the one hand, the aim of exploitation is to intensify the research around the most promising areas of the research space, which are in most cases close to the best-found solutions. On the other hand, it comes to diversifying the research by encouraging the exploration in order to discover new and better areas of the research space. The behavior of ants in relation to this duality between exploitation and exploration can be affected by the adjustment of the parameter values.

A comparative study was conducted on the hybrid ant colony algorithm MMACS. Firstly, this study is intended to present the behavior of MMACS algorithm and its dependencies on the values given to parameters while solving SCKP. Secondly, a comparison of performances of MMACS algorithm and two well-known Ant Colony Algorithms: the Max-Min Ant System (MMAS) and the Ant Colony System (ACS) was provided. Finally, MMACS algorithm was compared with two recent state of art algorithms that show significant results when solving the SCKP to optimality.

The paper has been organized as follows. In the next section, we define the Strongly Correlated Knapsack Problem. We present the studied Ant Colony Opti-

*Corresponding Author: Wiem Zouari, wiem.zouari@ensi-uma.tn

https://dx.doi.org/10.25046/aj030601

mization algorithms (ACO) in section 2. In section 3, we introduce the parameters in ACO. Then, we define the local search in section 4. Finally, experiments and results are presented in section 5.

2 Strongly Correlated Knapsack Problem (SCKP)

The SCKP problem is a NP-hard problem, whose goal is to find a subset of items that maximizes an objective function while satisfying resource constraints. In SCKP, the profit of each item is linearly related to its weight, in other words, the profit of an item is equal to its weights plus a fixed constant. The complexity of this problem compared to the classical knapsack problem resides in the strong correlation between the variables that characterize the problem. According to Pisinger in [2], the strongly correlated instances are hard to solve for two reasons. First, they are badly conditioned in the sense that there is a large gap between the continuous and integer solution of the problem. Then, sorting the items according to decreasing efficiencies corresponds to a sorting according to the weights. Thus, for any small interval of the ordered items (i.e. a "core"), there is a limited variation in the weights, making it difficult to equally satisfy the capacity constraint. SCKP can be formulated as follows:

$$\max \sum_{i=1}^{n} (w_i + k)x_i \tag{1}$$

subject to constraint:

$$\sum_{i=1}^{n} w_i x_i \leq c$$

$$x_i \in \{0,1\}, i = 1,...,n$$

where x_i is a decision variable associated with an item i, which has value 1 if the item is selected and 0 otherwise, w_i is the weight of the item i uniformly random $[1,R]$, k is a positive constant, c is the knapsack capacity and n is the number of items. The capacity of the knapsack c is proposed by Pisinger in [2] and it is obtained as follows :

$$c = \frac{i}{S+1} \sum_{j=1}^{n} w_j \tag{2}$$

where S is the series of instances as such, for each instance, a series of $S = 100$ instances is performed and $i = 1,..,S$ corresponds to the test instance number.
From equation (1), the profit $profit(x)$ of a solution x can be described as follows:

$$profit(x) = \sum_{i=1}^{n} w_i x_i + kb \tag{3}$$

where b is the number of items in x.
According to equation (3), the maximization of the $profit(x)$ means the maximization of the number of the selected items. In other words, items with the lowest weights should be first selected until the sum of weights is about to exceed the capacity c. This can be achieved through the use of greedy algorithms [3, 4]. However, greedy does not guarantee optimal solutions since it chooses the locally most attractive item with no concern for its effect on global solutions.
The convergence to local optima caused by greedy algorithms, called stagnation, should be avoided, hence the idea of alternation between greedy and stochastic approaches.
The proposition of Pisinger in [5] is one of the most well-known works that shows significant results when solving the SCKP to optimality. Pisinger proposed a specialized algorithm for this problem where he used a surrogate relaxation to transform the problem into a Subset-sum problem. He started the resolution by applying a greedy algorithm, then he used a 2-optimal heuristic and a dynamic programming algorithm to solve the problem to optimality. More recently, Han [6] proposed an evolutionary algorithm inspired by the concept of quantum computing. The study in [6] shows that the proposed algorithm, called Quantum-Inspired Evolutionary Algorithm (QEA), can find high quality results when solving the strongly correlated knapsack problems.

3 Ant Colony Optimization (ACO)

The ACO [7, 8] is a constructive population-based metaheuristic inspired from the real ants' behavior, seeking an adequate path between their colony and a food source, which is often the shortest path. The communication between ants is mediated by trails of a chemical substance called pheromone. Several ant colony optimization algorithms have been proposed in the literature. In this section, we present the Max-Min Ant System proposed by Stützle and Hoos [9, 10], the Ant Colony System proposed by Gambardella and Dorigo [11] and a recent hybrid ant colony algorithm called MMACS.

3.1 MMAS

The Max-Min Ant System [9, 10] is one of the most effective solvers of certain optimization problems. In MMAS, ants apply a random proportional rule to select the next item. The probabilistic action choice rule is defined as follows:

$$P_{ij}^{k} = \frac{[\tau_{ij}]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{l \in N_i^k}[\tau_{il}]^{\alpha}.[\eta_{il}]^{\beta}} \tag{4}$$

where τ and η are successively the pheromone factor and the heuristic factor, α and β are two parameters that determine the relative influence of the pheromone trail and the heuristic information and N_i^k is the feasible neighborhood of an ant k that selected an item i and chooses to select an item j. Besides, MMAS exploits the best solutions found by letting only the best

ant deposit pheromone. This best ant can be the one which found the best solution during the last iteration or the one which found the solution from the beginning of the execution. The pheromone update can be formulated as follows:

1. Pheromone evaporation applied to all components:

$$\tau_{ij} \leftarrow (1-\rho).\tau_i j \tag{5}$$

2. Pheromone update applied to components selected by the best ant:

$$\tau_{ij} \leftarrow \tau_{ij} + \Delta_{ij}^{bs} \tag{6}$$

Then, MMAS introduces bounds to limit the range of pheromone trails to $[\tau_{min}, \tau_{max}]$ in order to escape a stagnation that can be caused by an excessive growth of pheromone trails. These pheromone trails are initialized, at the beginning, to upper pheromone trail limit to ensure the exploration of the research space, and reinitialized when system approaches stagnation.

3.2 ACS

The ACS algorithm [11] achieves performance improvement through the use of a more aggressive action choice rule. In ACS, ants choose items according to an aggressive action choice rule called pseudorandom proportional rule given as follows:

$$s = \begin{cases} argmax_{l \in N_i^k} \{\tau_{il}.[\eta_{il}]^\beta\} & \text{if } q \leq q_0 \\ S & \text{otherwise} \end{cases} \tag{7}$$

where q is a random variable uniformly distributed in $[0,1]$, q_0 $(0 \leq q_0 \leq 1)$ and S is a random variable selected according to the probability distribution given in equation (4). Besides, only one ant called the best-so-far-ant is allowed to deposit pheromone after each iteration. Thus, the global pheromone trail update is given as follows:

$$\tau_{ij} \leftarrow (1-\rho).\tau_{ij} + \rho.\Delta_{ij}^{bs}(s) \tag{8}$$

This pheromone trail update is applied to only components in the best-so-far solution, where the parameter ρ represents pheromone evaporation. In addition to the global pheromone update, ants use a local pheromone update rule that is applied immediately after choosing a new item during the solution construction, given as follows:

$$\tau_{ij} \leftarrow (1-\epsilon).\tau_{ij} + \epsilon.\tau_0 \tag{9}$$

where $0 < \epsilon < 1$ and τ_0 are two parameters such that the τ_0 value is equal to the initial value of the pheromone trails which is 0.1. The local update happens during the solution construction in order to prevent other ants to make the same choices. This increases the exploration of alternative solutions.

3.3 MMACS

The MMACS algorithm combines Max-Min Ant System with Ant Colony System and an algorithm based on the 2-opt heuristic. In fact, the scheme of MMACS is based on ACO scheme presented in [12] and ACS scheme presented in [11] in a way that it uses an MMAS pheromone update rule and a choice rule inspired from the ACS aggressive action choice rule. In MMACS, the minimum and the maximum pheromone amounts are limited to an interval $[\tau_{min}, \tau_{max}]$, like MMAS, in order to avoid premature stagnation. Initially, the pheromone trails are set to τ_{max}. After the construction of all solutions in one cycle, the best ant updates the pheromone trails by applying a rule similar to MMAS pheromone update rule. Indeed, once all ants finish the solutions construction, the pheromone trails are decreased to simulate evaporation by multiplying each component by a pheromone persistence ratio equal to $(1-\rho)$ where $0 \leq \rho \leq 1$ as given by equation (5). After that, an amount of pheromone is laid on the best solution found by ants by applying the pheromone update rule given in equation (6), where the amount of pheromone trails is calculated as follows:

$$\Delta\tau_{S_k} = \frac{1}{1 + profit(S_{best}) - profit(S_k)} \tag{10}$$

S_{best} represents the best solution built since the beginning and S_k is the best solution of a cycle.
Besides, in MMACS, each ant constructs a solution by applying the choice rule, where the decision making is based on both:

1. A random proportional rule that selects a random item using the probability distribution.
2. A guided selection rule that chooses the next item as the best available option.

Like ACS, MMACS balances between greedy and stochastic approaches by applying the pseudorandom proportional rule (7). Actually, at each construction step, an ant k chooses a random variable q uniformly distributed in $[0,1]$. If q is less than a fixed parameter q_0 such as $0 \leq q_0 \leq 1$, the ant makes the best possible choice as indicated by the pheromone trails and the heuristic information (exploitation) else, with a probability $1-q_0$, the ant applies the random proportional rule (4) to select the next item (biased exploration).
The heuristic factor used in the probability rule (4) is given as follows:

$$\eta_{S_k}(o_j) = \frac{d_{S_k}}{w_j} \tag{11}$$

where d_{S_k} is the remaining capacity when an ant k built a solution S_k and it is given as follows:

$$d_{S_k} = c - \sum_{g \in S_k} w_g \tag{12}$$

As shown in equation (11), the heuristic information value and the item weight are inversely proportional. Consequently, the more the weight value decreases the

more the heuristic information value increases.
Added to that, the closer the remaining capacity and the item weight are, the more the heuristic information value increases. This can be helpful at the end of the execution when the knapsack is about to be filled.
At last, the execution of MMACS algorithm ends either when an optimum is found, or in the worst cases, it ends after a fixed number of iterations. The pseudo-code of MMACS algorithm is represented by algorithm 1.

Algorithm 1 MMACS pseudo-code applied to SCKP

```
Initialize pheromone trails to τ_max
repeat
  repeat
    Construct a solution
    Update S_best
  until maximum number of ants is reached or optimum is found
  Update pheromone trails
until maximum number of cycles is reached or optimum is found
Apply a local search algorithm
```

S_{best} is the best solution found all along the execution.

The construction procedure can be represented by algorithm 2.

Algorithm 2 Construct Solution

```
Select randomly a first item
Remove from candidates each item that violates resource constraints
while Candidates ≠ ∅ do
  if a randomly chosen q is greater than q_0 then
    Choose item o_j from Candidates with probability P_ij^k
  else
    Choose the next best item
  end if
  Remove from candidates each item that violates resource constraints
end while
```

4 Parameters in ACO

In the ACO algorithm, relevant parameters that request reasonable settings are the heuristic information parameter β, the pheromone parameter α and the pheromone evaporation rate ρ. Those parameters can influence the algorithm performance by improving its convergence speed and its global optimization ability.

4.1 The heuristic information parameter β

The ants' solution construction is biased by a heuristic value η that represents the attractiveness of each item. The parameter β determines the relative importance of this heuristic value. In our case, the heuristic information makes items characterized by little weights as desirable choices. In other words, the increase in the β value can be triggered by the selection of items which have little weights. This behavior is close to that of greedy algorithm. However, the decrease in the value of β makes the heuristic factor unprofitable. As a result, ants fall easily in local optima.

4.2 The pheromone parameter α

Besides heuristic value, the ants' solution construction is influenced by the pheromone trails. The pheromone parameter α determines the relative influence of the pheromone trails τ. Indeed, the parameter α reflects the importance of the amplification of pheromone amounts. In other words, the increase of α favors the choice of items associated with uppermost pheromone trails values. In case the value of α is considerable, ants tend to choose the same solution components. This behavior is caused by the strong cooperation between them so ants drift towards the same part of the search space. In such case, the convergence speed of algorithm accelerates and consequently, it causes the fall in local optima. This gives rise to the need to prevent this premature convergence. Then, several tests were conducted to evaluate the influence of α on solutions' quality. Additional experiments were carried out to examine the similarity of solutions in one cycle. The analysis of the similitude of solutions allows appropriately assigning the value of α in order to avoid the premature stagnation of the search that can be caused by the excessive reliance upon pheromone trails at the expense of the heuristic information.

4.3 The pheromone evaporation rate ρ

The amount of pheromone decreases to simulate evaporation by multiplying each component by a constant evaporation ratio equal to $1-\rho$. This pheromone trails reduction gives ants the possibility of abandoning bad decisions previously taken. In fact, the pheromone value of an unchosen item decreases exponentially with the number of iterations.

5 Local search

Local search algorithms are usually used in most applications of ACO to combinatorial optimization problems in order to improve solutions found by ants. Among those algorithms, we cite 2-opt heuristic. The 2-opt [13] is a simple local search algorithm. When applied to knapsack problems, it consists of exchanging an item present in the current solution with another that is not part of this solution in order to improve it. The new solution should satisfy constraints and it would be better or equal to the old one. In other words, the 2-opt algorithm takes a current solution as input and returns a better accepted solution to the problem,

Table 1: Default parameter settings for MMAS, ACS and MMACS algorithms

ACO algorithm	α	β	ρ	$ants$	$cycles$	q_0
MMAS	1	2	0.02	n	20	_
MMAS + 2-opt	1	2	0.2	25	20	_
ACS	_	2	0.1	10	20	0.9
ACS + 2-opt	_	2	0.1	10	20	0.98
MMACS	1	5	0.02	20	20	0.9

if it exists. The 2-opt algorithm is used once the ants have completed their solution construction, thereby improving the solution by approaching the best one or even reaching it. Our proposed 2-opt algorithm can be written as represented by algorithm 3.

Algorithm 3 A 2–opt pseudo-code applied to SCKP

```
Initialize Candidates by observing S_best
repeat
  for each item o_j ∈ Candidates do
    for each item o_i ∈ S_best do
      S'_best = Swap (o_i, o_j)
      if constraints are satisfied by S'_best and S'_best is
      better than S_best then
        Update best solution
      end if
    end for
  end for
until no improvement is made
```

6 Computational results

In this section, we study the results of a set of experiments that was carried out to determine the efficacy of the MMACS algorithm. The proposed algorithm was programmed in C++, and compiled with GNU g++ on an Intel Core i7-4770 CPU processor (3.40 GHz and 3.8 GB RAM).
Through the experimentations, we analyze the influence of the parameters' selection on the MMACS performances. Then, we identify the convenient parameter settings that produce better results. Those parameter settings are employed for the rest of the experiments. At a later stage, we compare the results of MMACS with those of the two well-known ant colony algorithms: MMAS [9, 10] and ACS [11]. After that, the results of MMACS are compared with those of the evolutionary algorithm QEA in [6] and to the optimal values.

6.1 Benchmark instances

In order to evaluate the performance of the MMACS algorithm, experiments were conducted on two sets of instances.

6.1.1 Pisinger Set

The first set contains 100 different instances with n items, where the number of items n varies from 50 to 2000. Those benchmark instances used in comparison with algorithms in [5] and [6] are available at the website (http://www.diku.dk/pisinger/codes.html).

6.1.2 Generated Set

The second set regroups 3 different instances having the number of items equal to 100, 250 and 500, respectively. Those instances used in comparison with the proposed algorithm in [6], were randomly generated using a generator similar to the one in [2].

6.2 Parametric analysis of MMACS

In order to evaluate the influence of parameters' values on MMACS performances, we conducted tests for different values of parameters and compared the obtained results. The experiments were realized on the Pisinger set instances of size 50, 100, 200 and 500, and for each instance, we applied 10 runs (10 runs * 100 instances * 4 knapsack problems). In each of these experiments, we fixed the parameters to their default values and we made the variation of the studied parameter. In fact, the MMAS and ACS default parameter settings were recommended by the authors in [14]. The default parameter settings are given in Table 1 and the various values for each parameter are presented in Table 2. Tables 3- 29 report the results of MMACS algorithm in response to the variation of the parameters. N is the number of items and R is the range of coefficients. Then, the presentation of the data is visualized using different curves. Figures 1- 5 present the effect of the studied parameters on MMACS performances. The abscissa axis of curves presented in those figures shows the instances' size that varies between 50 and 500. In each figure, left and right plots show the behavior of MMACS algorithm while solving the SCKP instances having the range of coefficients equal to 1000 and 10000, respectively. In those curves, we examine the percentage of exact solutions, the relative deviation of the best solution found by MMACS from the optimal solution value and the execution time in terms of the studied parameter.

Table 2: Parameter settings used in experiments for MMACS algorithm

Control parameter	**Value**
α	1, 2, 3, 4, 5
β	1, 2, 3, 4, 5
ρ	0.01, 0.02, 0.4, 0.5, 0.8, 1
q_0	0, 0.5, 0.75, 0.9, 0.99
m	1, 5, 10, 20, 50, 1000

6.2.1 Influence of parameter α

We set the value of β to 5 and ρ to 0.02. After that, we make the change of the value of α in order to study its influence on the solutions' quality and the execution time. Tables 3- 6 represent the results after applying MMACS to SCKP where α varies between 1 and 5. Results are visualized using curves in Figure 1. In fact, the differences among the various settings of α are almost insignificant. However, the value of α equal to 1 gives better results in terms of the most reduced execution time. Additional experiments are conducted to analyze the similitude of solutions in one cycle in order to appropriately assign the value of α. In fact, we propose to calculate a similarity ratio proposed in [15]. The similarity ratio associated with a set of solutions S is defined as follows:

$$ratio = \frac{\sum_{v_i \in V}(freq[i].(freq[i]-1))}{(|S|-1).\sum_{S_k \in S}|S_k|} \qquad (13)$$

where V is a set of items and $freq[i]$ is the frequency of the object v_i in the solutions of S. Thus, this ratio is equal to 1 if all the solutions of S are identical and it is equal to 0 if the intersection of solutions of S is empty. In those experiments, the number of cycles varies from 10 to 25. For each number of cycles, each ant's solution was compared with others. Experiments are conducted on the Pisinger set instances of size 100 and for both range of coefficients 1000 and 10000. Figure 2 shows the influence of the pheromone trails on the similarity of the solutions built during the execution of the MMACS algorithm, and for different values of the α parameter that determine the influence of the pheromone on the behavior of the ants. The curves show that the solutions are remarkably similar where the similarity ratio varies between the values 0.6 and 0.8 beginning with the number of cycles at about 15. In other words, ants are not deeply influenced by pheromone trails and they are obviously focusing in a short time on a very small area of the research space that they explore intensely.

6.2.2 Influence of parameters β and ρ

In this section, we examine the influence of different values of the heuristic information parameter β and the pheromone evaporation rate ρ. In this context, we fix the value of α to 1 then we make a simultaneous variation of the two variables β and ρ. We modify the values of β in order to control the influence of the heuristic information and to examine its effect on the MMACS performances. Besides, we make the variety of ρ values in order to study its influence on the items' selection and consequently on the MMACS performances. Tables 7- 21 present results after applying MMACS to SCKP where β varies between 2 and 5 and ρ values are 0.01, 0.02, 0.4, 0.5, 0.8 and 1.

Results of MMACS algorithms with the fixed parameter β are almost similar to all values of ρ.

The MMACS algorithms with the fixed parameter $\beta 1$ and $\beta 2$ work in a similar way. In fact, the results show that the percentages of exact solutions decrease considerably in terms of number of items for both ranges of coefficients 1000 and 10000. Consequently the values of gap increases. Besides, the execution time results vary in the same way for the six values of ρ.

The MMACS algorithms with the fixed parameter $\beta 3$ and $\beta 4$ behave in a similar way for almost all problems with both ranges of coefficients 1000 and 10000. Results show that the percentages of exact solutions for both algorithms increase for number of items between 50 and 200. Then these percentages decrease considerably for the large number of items 500. Consequently, the gap results progress in a reverse way.

However, the MMACS algorithm with the fixed heuristic parameter $\beta 5$ shows acceptable results. Results are presented in Figure 3, each curve represents a value of ρ. The curves have almost the same evolutions with insignificant differences between the values. In fact, the percentage of exact solution presents an increase in terms of number of items. Besides, the gap results show a remarkable decrease for large instances. Then, the execution time has the same variation for all values of ρ.

However, the value of $\rho 2$ equal to 0.02 selected by MMACS can be modified to $\rho 1$ in order to make a slight improvement in gap values for large instances with a range of coefficients equal to 10000.

6.2.3 Influence of parameter q_0

In MMACS, we study the effect of q_0 that represents the probability of selecting the best available choices in equation 7. Results are given in Tables 22- 25 and presented in curves of Figure 4. The curves in Figure 5 are associated with the values of q_0: 0, 0.5, 0.75, 0.9 and 0.99. For all instances, among those values only 0.9 and 0.99 show an increase in terms of the percentage of the exact solutions. As regards the other values of q_0 MMACS does not succeed in finding, in most cases, the exact solution. This is clearly represented by its decreasing curves. As to the execution time, the five curves are growing in almost the same way. However, the curve that corresponds to q_0 value equal to 0.9 reaches the lowest values. For large instances, the differences between the curves are significant. The best results that correspond to the lowest gap values is represented by the curve associated to the value of q_0 equal to 0.9. We conclude that MMACS has the same behavior as ACS regarding the q_0 parameter, where the values close to 1 present the good ones as suggested in the literature.

6.2.4 Influence of colony size

Tables 26- 29 show the effect of colony size on the quality of the solutions. In these tables, the ant colony size m varies between 1 and 100. Results are compared in Figure 5. As shown by curves in Figure 5, the increase in the ants' number improves the percentage of exact solutions. This increase causes the growth of the execution time, although in practice, we generally seek to

Table 3: Percentage of exact solutions found by MMACS. The value of α varies between 1 and 5.

N	R	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
50	1000	82.0%	79.0%	83.0%	77.0%	80.0%
	10000	52.0%	48.0%	51.0%	48.0%	46.0%
100	1000	90.0%	91.0%	93.0%	88.0%	91.0%
	10000	63.0%	60.0%	58.0%	65.0%	59.0%
200	1000	92.0%	94.0%	92.0%	90.0%	92.0%
	10000	79.0%	82.0%	82.0%	80.0%	84.0%
500	1000	98.0%	97.0%	97.0%	9.0%	93.0%
	10000	79.0%	78.0%	76.0%	76.0%	81.0%

Table 4: Percentage of perfect solutions found by MMACS. The value of α varies between 1 and 5.

N	R	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
50	1000	52.0%	50.0%	54.0%	48.0%	52.0%
	10000	13.0%	12.0%	15.0%	12.0%	10.0%
100	1000	82.0%	84.0%	86.0%	81.0%	83.0%
	10000	46.0%	44.0%	41.0%	48.0%	41.0%
200	1000	88.0%	87.0%	88.0%	86.0%	87.0%
	10000	72.0%	75.0%	74.0%	73.0%	78.0%
500	1000	96.0%	94.0%	95.0%	96.0%	92.0%
	10000	78.0%	76.0%	75.0%	74.0%	80.0%

Table 5: Averages of solutions found by MMACS. The value of α varies between 1 and 5.

N	R	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
50	1000	0.02934	0.05073	0.06385	0.02732	0.03439
	10000	0.02485	0.02534	0.02009	0.02104	0.02113
100	1000	0.00887	0.03124	0.01363	0.01113	0.01537
	10000	0.00450	0.00445	0.00534	0.00668	0.00497
200	1000	0.00248	0.00239	0.00149	0.00213	0.00198
	10000	0.00087	0.00143	0.00102	0.00154	0.00125
500	1000	0.00059	0.00102	0.00056	0.00220	0.00445
	10000	0.00020	0.00049	0.00064	0.00038	0.00312

Table 6: Execution time of MMACS. The value of α varies between 1 and 5.

N	R	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$
50	1000	0.05370	0.08514	0.07676	0.08809	0.08404
	10000	0.11360	0.20702	0.14372	0.14228	0.15597
100	1000	0.36749	0.35563	0.49085	0.34433	0.36030
	10000	0.73181	1.13194	1.67862	0.84895	0.97173
200	1000	1.64040	3.23461	2.36788	2.40549	2.40237
	10000	4.02055	7.20106	8.16290	4.69110	7.43489
500	1000	10.5397	48.9903	46.4050	37.3577	28.8331
	10000	38.1680	71.5076	64.0864	72.4135	59.0820

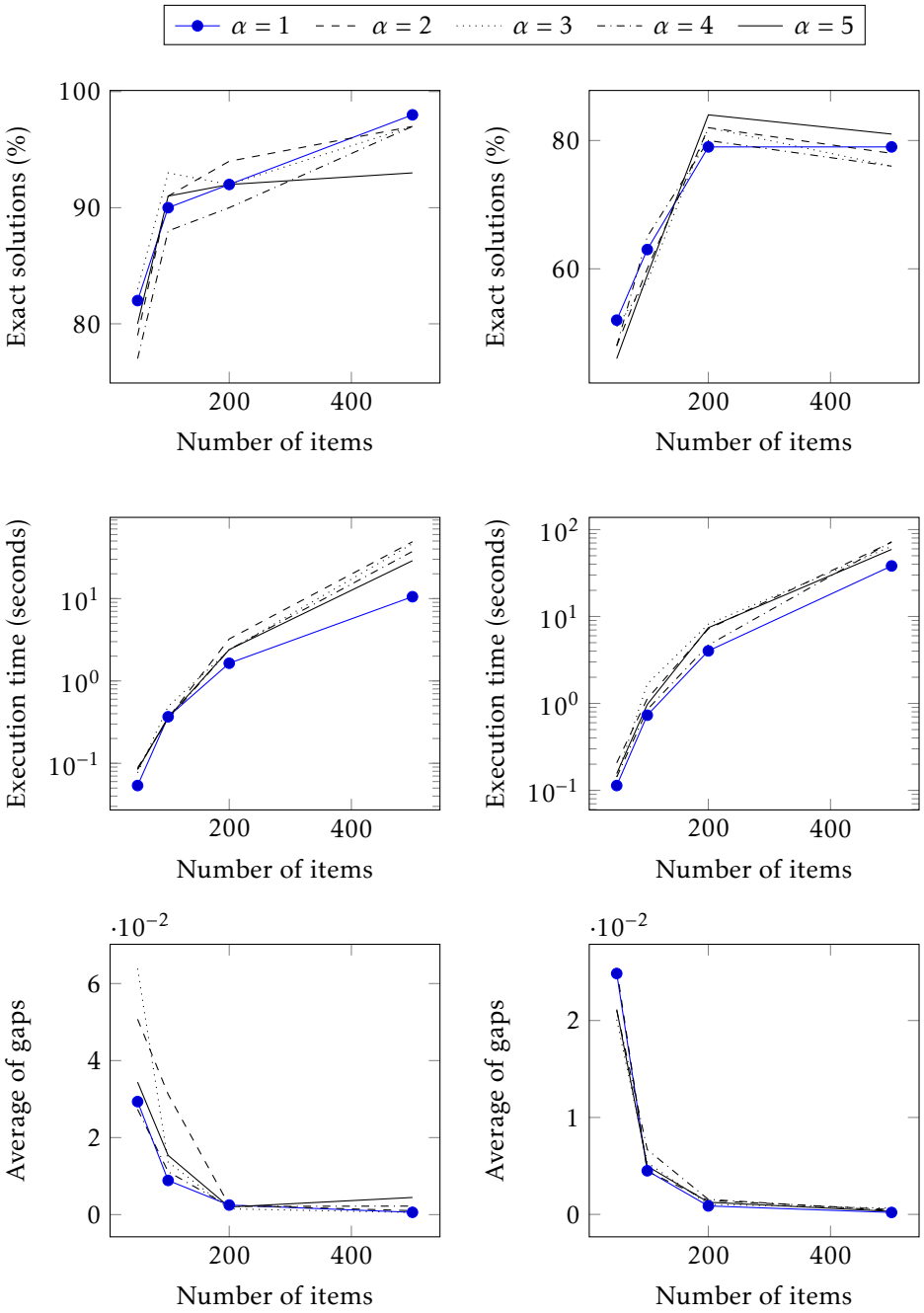


Figure 1: MMACS with various values of α. Plots on the left show results for SCKP instances with a range of coefficients equal to 1000 and plots on the right show results for SCKP instances with a range of coefficients equal to 10000.

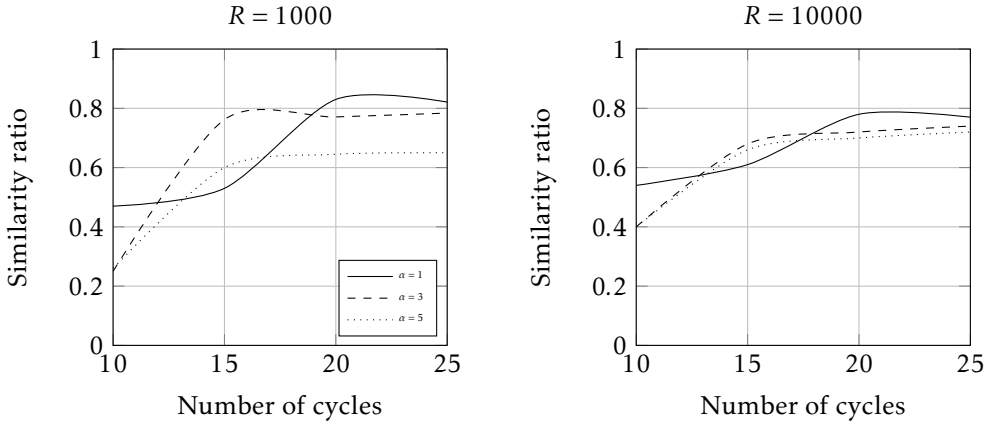


Figure 2: Influence of α on the similarity of solutions found by MMACS algorithm for SCKP instances with 100 items.

Table 7: Percentage of exact solutions found by MMACS. The heuristic information value is fixed to $\beta 1$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 1$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	82.0%	81.0%	83.0%	84.0%	83.0%	84.0%
	10000	55.0%	54.0%	48.0%	55.0%	50.0%	57.0%
100	1000	91.0%	86.0%	90.0%	87.0%	88.0%	87.0%
	10000	53.0%	51.0%	57.0%	53.0%	54.0%	55.0%
200	1000	84.0%	84.0%	82.0%	82.0%	84.0%	81.0%
	10000	55.0%	49.0%	46.0%	52.0%	53.0%	51.0%
500	1000	35.0%	40.0%	41.0%	37.0%	37.0%	34.0%
	10000	24.0%	28.0%	23.0%	27.0%	22.0%	27.0%

Table 8: Averages of solutions found by MMACS. The heuristic information value is fixed to $\beta 1$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 1$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.08984	0.02556	0.03288	0.03032	0.03363	0.03963
	10000	0.02220	0.02653	0.01738	0.02172	0.01726	0.01966
100	1000	0.02126	0.01814	0.01206	0.02616	0.01718	0.04464
	10000	0.00870	0.00930	0.01639	0.00883	0.00897	0.01018
200	1000	0.00832	0.04333	0.02262	0.04705	0.00998	0.02377
	10000	0.00676	0.01265	0.00538	0.00577	0.01255	0.01741
500	1000	0.00676	0.16979	0.15236	0.11632	0.15827	0.17096
	10000	0.11085	0.12811	0.13676	0.12798	0.11106	0.12794

Table 9: Execution time of MMACS. The heuristic information value is fixed to $\beta 1$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 1$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.05492	0.05765	0.06821	0.05596	0.05465	0.05635
	10000	0.05567	0.05591	0.05756	0.05331	0.05382	0.05712
100	1000	0.39064	0.38973	0.76695	0.38287	0.39028	0.39951
	10000	0.38998	0.38882	0.77140	0.38358	0.39128	0.39009
200	1000	2.98575	3.00137	2.94311	2.93495	2.91098	2.95657
	10000	3.00461	2.95496	2.93763	2.91634	2.91860	2.95932
500	1000	48.0960	47.6133	46.3170	54.1633	46.7279	47.2380
	10000	47.9146	46.8829	46.5491	46.5299	46.7395	47.3001

Table 10: Percentage of exact solutions found by MMACS. The heuristic information value is fixed to $\beta 2$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 2$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	80.0%	84.0%	82.0%	81.0%	84.0%	82.0%
	10000	52.0%	47.0%	51.0%	55.0%	54.0%	52.0%
100	1000	91.0%	93.0%	91.0%	90.0%	90.0%	91.0%
	10000	59.0%	59.0%	66.0%	60.0%	60.0%	55.0%
200	1000	86.0%	92.0%	90.0%	89.0%	90.0%	93.0%
	10000	62.0%	67.0%	68.0%	69.0%	65.0%	65.0%
500	1000	60.0%	63.0%	52.0%	58.0%	44.0%	59.0%
	10000	41.0%	37.0%	36.0%	36.0%	36.0%	33.0%

Table 11: Averages of solutions found by MMACS. The heuristic information value is fixed to $\beta 2$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 2$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.03299	0.03350	0.03077	0.05647	0.05426	0.03865
	10000	0.01685	0.01709	0.01783	0.01535	0.02255	0.02233
100	1000	0.01535	0.02211	0.01249	0.01724	0.01435	0.01480
	10000	0.00785	0.00968	0.01008	0.00698	0.00798	0.01020
200	1000	0.00337	0.00443	0.00610	0.00417	0.00520	0.00423
	10000	0.00116	0.00208	0.00217	0.00180	0.00181	0.00275
500	1000	0.06840	0.08212	0.06738	0.06309	0.07076	0.06730
	10000	0.03134	0.05077	0.03474	0.03055	0.03297	0.02785

Table 12: Execution time of MMACS. The heuristic information value is fixed to $\beta 2$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 2$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.05313	0.05536	0.05604	0.05722	0.05452	0.05769
	10000	0.05358	0.05588	0.05365	0.05522	0.05618	0.05340
100	1000	0.37927	0.39185	0.38238	0.38306	0.38370	0.38917
	10000	0.38958	0.38976	0.39358	0.38518	0.38692	0.38878
200	1000	2.91574	2.96851	2.92166	2.91118	2.96750	2.97729
	10000	2.93331	2.91021	2.91412	2.95463	2.92098	2.94065
500	1000	46.3217	68.3414	46.0147	48.9379	47.6850	48.1014
	10000	46.0396	45.7421	46.2623	48.0479	46.4917	54.1009

Table 13: Percentage of exact solutions found by MMACS. The heuristic information value is fixed to $\beta 3$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 3$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	81.0%	84.0%	83.0%	83.0%	79.0%	85.0%
	10000	50.0%	48.0%	49.0%	50.0%	54.0%	53.0%
100	1000	95.0%	89.0%	89.0%	91.0%	90.0%	94.0%
	10000	59.0%	63.0%	62.0%	61.0%	56.0%	57.0%
200	1000	92.0%	89.0%	93.0%	94.0%	92.0%	93.0%
	10000	72.0%	80.0%	77.0%	72.0%	73.0%	74.0%
500	1000	84.0%	88.0%	85.0%	84.0%	80.0%	86.0%
	10000	52.0%	50.0%	53.0%	46.0%	46.0%	49.0%

Table 14: Averages of solutions found by MMACS. The heuristic information value is fixed to $\beta 3$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 3$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.02858	0.03836	0.03311	0.03606	0.05618	0.03919
	10000	0.02232	0.02145	0.02609	0.02073	0.02204	0.02508
100	1000	0.02088	0.01689	0.01302	0.01218	0.01536	0.02067
	10000	0.00652	0.00897	0.00696	0.00931	0.00977	0.01037
200	1000	0.00353	0.00350	0.00293	0.00233	0.00261	0.00330
	10000	0.00093	0.00095	0.00117	0.00118	0.00161	0.00128
500	1000	0.02795	0.05289	0.04132	0.03866	0.03520	0.03838
	10000	0.01185	0.01029	0.01414	0.01199	0.00983	0.00629

Table 15: Execution time of MMACS. The heuristic information value is fixed to $\beta 3$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 3$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.07581	0.07366	0.07465	0.07367	0.07425	0.07242
	10000	0.07422	0.07476	0.07528	0.07497	0.08080	0.07237
100	1000	0.46370	0.46119	0.46399	0.46445	0.45845	0.45893
	10000	0.46346	0.45978	0.46745	0.46602	0.46544	0.46089
200	1000	3.27010	3.26748	3.29506	3.30715	3.28653	3.24697
	10000	3.26838	3.27883	3.28464	3.30160	3.27514	3.24717
500	1000	50.0693	49.6441	49.2067	48.8005	48.5047	48.1483
	10000	49.8877	48.5299	48.9702	51.1495	48.8320	48.0786

Table 16: Percentage of exact solutions found by MMACS. The heuristic information value is fixed to $\beta 4$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 4$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	84.0%	84.0%	82.0%	81.0%	80.0%	84.0%
	10000	51.0%	54.0%	51.0%	51.0%	51.0%	55.0%
100	1000	89.0%	90.0%	91.0%	91.0%	90.0%	93.0%
	10000	59.0%	61.0%	62.0%	67.0%	60.0%	53.0%
200	1000	91.0%	93.0%	92.0%	93.0%	91.0%	94.0%
	10000	83.0%	80.0%	81.0%	83.0%	78.0%	85.0%
500	1000	90.0%	93.0%	92.0%	89.0%	94.0%	85.0%
	10000	70.0%	70.0%	77.0%	65.0%	71.0%	74.0%

Table 17: Averages of solutions found by MMACS. The heuristic information value is fixed to $\beta 4$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 4$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.03218	0.03467	0.03252	0.03324	0.03187	0.03196
	10000	0.01684	0.02137	0.02540	0.02021	0.02979	0.02631
100	1000	0.01555	0.02284	0.01867	0.02045	0.02287	0.01441
	10000	0.00541	0.01040	0.00831	0.00895	0.00965	0.00892
200	1000	0.00263	0.00192	0.00176	0.00350	0.00185	0.00218
	10000	0.00104	0.00155	0.00142	0.00118	0.00127	0.00097
500	1000	0.02596	0.02232	0.03127	0.01763	0.01168	0.01989
	10000	0.00606	0.00439	0.00595	0.01001	0.00683	0.01049

Table 18: Execution time of MMACS. The heuristic information value is fixed to $\beta 4$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 4$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.07929	0.08029	0.07424	0.07557	0.10963	0.07293
	10000	0.07846	0.07948	0.07611	0.07561	0.11752	0.07161
100	1000	0.49581	0.48432	0.48316	0.48317	0.48473	0.45665
	10000	0.49517	0.47001	0.49464	0.48414	0.48985	0.45644
200	1000	3.38250	3.28004	3.46215	3.45131	3.09758	3.22518
	10000	3.51687	3.29262	3.30075	3.45131	3.46598	3.23389
500	1000	48.3098	48.9813	47.9923	47.9734	49.1355	47.8999
	10000	48.7001	48.5943	48.7463	48.2574	48.2564	47.7792

Table 19: Percentage of exact solutions found by MMACS. The heuristic information value is fixed to $\beta 5$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 5$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	82.0%	81.0%	81.0%	83.0%	79.0%	82.0%
	10000	49.0%	50.0%	47.0%	51.0%	44.0%	51.0%
100	1000	89.0%	91.0%	89.0%	88.0%	90.0%	89.0%
	10000	55.0%	58.0%	61.0%	55.0%	53.0%	59.0%
200	1000	91.0%	92.0%	90.0%	93.0%	92.0%	91.0%
	10000	79.0%	76.0%	81.0%	79.0%	76.0%	81.0%
500	1000	97.0%	98.0%	96.0%	95.0%	98.0%	97.0%
	10000	80.0%	78.0%	79.0%	79.0%	81.0%	77.0%

Table 20: Averages of solutions found by MMACS. The heuristic information value is fixed to $\beta 5$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 5$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.03247	0.03550	0.04044	0.06754	0.03757	0.03616
	10000	0.01910	0.03156	0.01838	0.01617	0.02353	0.02000
100	1000	0.01684	0.01483	0.01291	0.01588	0.01642	0.01395
	10000	0.00378	0.00786	0.00950	0.00682	0.00966	0.00802
200	1000	0.00218	0.00241	0.00284	0.00188	0.00210	0.00213
	10000	0.00129	0.00101	0.00097	0.00116	0.00158	0.00061
500	1000	0.00058	0.00188	0.00050	0.00265	0.00097	0.00283
	10000	0.00073	0.00460	0.00055	0.00069	0.00033	0.00348

Table 21: Execution time of MMACS. The heuristic information value is fixed to $\beta 5$ and the value of ρ varies between 0.01 and 1.

N	R	$\beta 5$					
		$\rho 1$	$\rho 2$	$\rho 3$	$\rho 4$	$\rho 5$	$\rho 6$
50	1000	0.07230	0.06982	0.07270	0.07362	0.07413	0.07302
	10000	0.07589	0.06936	0.07335	0.07472	0.07412	0.07157
100	1000	0.46598	0.47126	0.43751	0.46137	0.45676	0.45724
	10000	0.47470	0.46507	0.46672	0.46696	0.45895	0.45802
200	1000	3.34288	3.31321	3.31792	3.29320	3.27194	3.24836
	10000	3.09774	3.33303	3.35658	3.33049	3.28618	3.23154
500	1000	53.2248	48.9947	49.9506	48.2442	45.1045	44.6440
	10000	48.7456	49.1683	49.5907	49.3678	45.7457	48.2131

Table 22: Percentage of exact solutions found by MMACS. The values of q_0 are 0, 0.5, 0.75, 0.9 and 0.99.

N	R	$q1$	$q2$	$q3$	$q4$	$q5$
50	1000	83.0%	93.0%	84.0%	82.0%	66.0%
	10000	46.0%	61.0%	53.0%	52.0%	39.0%
100	1000	50.0%	86.0%	94.0%	90.0%	77.0%
	10000	11.0%	53.0%	67.0%	63.0%	36.0%
200	1000	15.0%	59.0%	88.0%	92.0%	88.0%
	10000	05.0%	23.0%	70.0%	79.0%	62.0%
500	1000	02.0%	12.0%	56.0%	98.0%	97.0%
	10000	01.0%	06.0%	24.0%	79.0%	85.0%

Table 23: Percentage of perfect solutions found by MMACS. The values of q_0 are 0, 0.5, 0.75, 0.9 and 0.99.

N	R	$q1$	$q2$	$q3$	$q4$	$q5$
50	1000	52.0%	62.0%	54.0%	52.0%	39.0%
	10000	12.0%	19.0%	16.0%	13.0%	6.0%
100	1000	44.0%	80.0%	87.0%	82.0%	68.0%
	10000	06.0%	36.0%	49.0%	46.0%	22.0%
200	1000	13.0%	56.0%	85.0%	88.0%	82.0%
	10000	05.0%	22.0%	66.0%	72.0%	56.0%
500	1000	01.0%	11.0%	55.0%	96.0%	95.0%
	10000	01.0%	06.0%	24.0%	78.0%	84.0%

Table 24: Averages of solutions found by MMACS. The values of q_0 are 0, 0.5, 0.75, 0.9 and 0.99.

N	R	$q1$	$q2$	$q3$	$q4$	$q5$
50	1000	0.02926	0.04270	0.07026	0.02934	0.31059
	10000	0.02430	0.01672	0.01380	0.02485	0.02395
100	1000	0.08169	0.03278	0.01605	0.00887	0.01448
	10000	0.04577	0.00236	0.00286	0.00450	0.00636
200	1000	0.16829	0.08070	0.00426	0.00248	0.00327
	10000	0.15506	0.03231	0.00146	0.00087	0.00090
500	1000	0.33937	0.13807	0.06764	0.00059	0.00447
	10000	0.29663	0.12393	0.13807	0.00020	0.00026

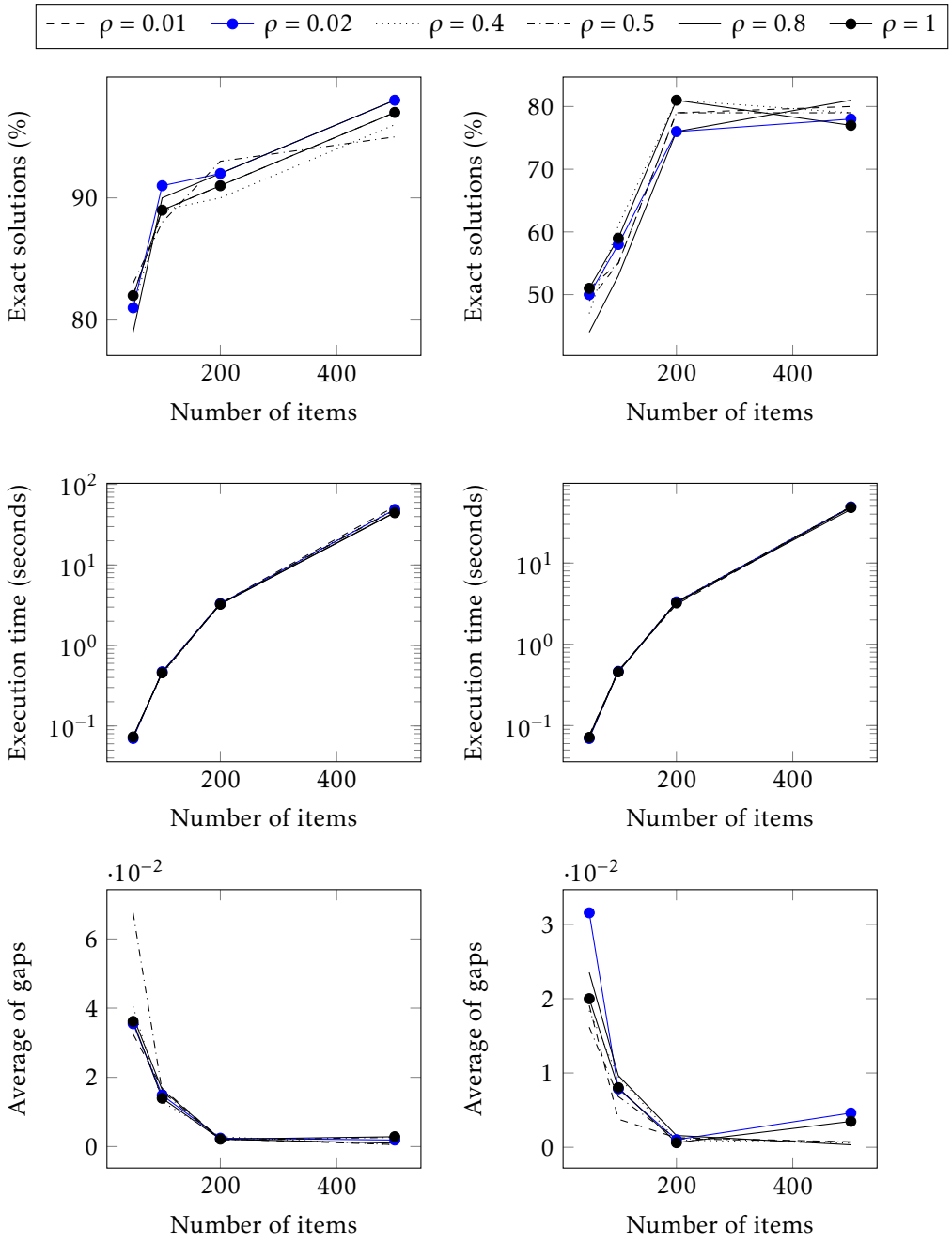


Figure 3: MMACS with a fixed heuristic parameter $\beta 5$ and various values of ρ . Plots on the left show results for SCKP instances with a range of coefficients equal to 1000 and plots on the right show results for SCKP instances with a range of coefficients equal to 10000.

Table 25: Execution time of MMACS. The values of q_0 are 0, 0.5, 0.75, 0.9 and 0.99.

N	R	$q1$	$q2$	$q3$	$q4$	$q5$
50	1000	0.11231	0.06590	0.07087	0.05370	0.17519
	10000	0.14675	0.13272	0.24615	0.11360	0.15682
100	1000	1.11215	0.45967	0.59802	0.36749	1.04415
	10000	1.40685	1.06039	1.16371	0.73181	2.03097
200	1000	11.0831	7.03010	6.07585	1.64040	4.58164
	10000	11.3605	9.8870	11.0704	4.02055	11.2200
500	1000	189.264	162.474	114.984	10.5397	14.0793
	10000	179.640	175.398	162.474	38.1680	47.3076

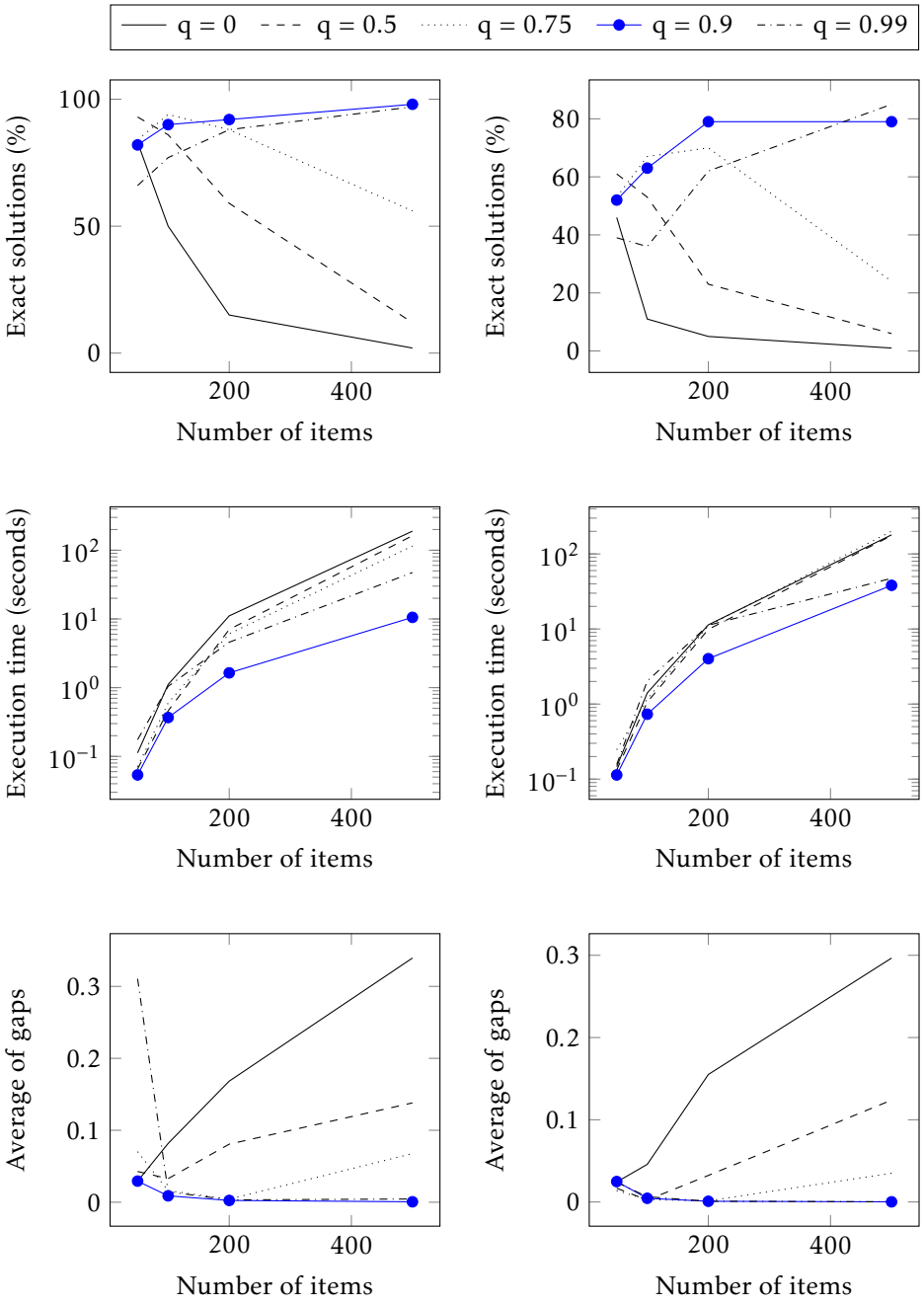


Figure 4: MMACS with various values of q_0. Plots on the left show results for SCKP instances with a range of coefficients equal to 1000 and plots on the right show results for SCKP instances with a range of coefficients equal to 10000.

Table 26: Percentage of exact solutions found by MMACS. The number of ants varies between 1 and 100.

N	R	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
50	1000	41.0%	67.0%	77.0%	82.0%	86.0%	89.0%
	10000	25.0%	39.0%	45.0%	52.0%	59.0%	66.0%
100	1000	56.0%	84.0%	84.0%	90.0%	95.0%	93.0%
	10000	20.0%	42.0%	48.0%	63.0%	73.0%	84.0%
200	1000	72.0%	88.0%	90.0%	92.0%	94.0%	93.0%
	10000	18.0%	52.0%	67.0%	79.0%	89.0%	89.0%
500	1000	68.0%	86.0%	93.0%	98.0%	98.0%	98.0%
	10000	19.0%	53.0%	72.0%	79.0%	87.0%	90.0%

Table 27: Percentage of perfect solutions found by MMACS. The number of ants varies between 1 and 100.

N	R	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
50	1000	18.0%	40.0%	48.0%	52.0%	57.0%	59.0%
	10000	2.0%	5.0%	10.0%	13.0%	20.0%	24.0%
100	1000	48.0%	75.0%	77.0%	82.0%	87.0%	86.0%
	10000	7.0%	29.0%	29.0%	46.0%	54.0%	62.0%
200	1000	63.0%	83.0%	85.0%	88.0%	88.0%	88.0%
	10000	13.0%	47.0%	62.0%	72.0%	82.0%	81.0%
500	1000	66.0%	84.0%	90.0%	96.0%	97.0%	97.0%
	10000	16.0%	52.0%	71.0%	78.0%	86.0%	89.0%

Table 28: Averages of solutions found by MMACS. The number of ants varies between 1 and 100.

N	R	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
50	1000	0.17092	0.04399	0.05144	0.02934	0.03614	0.04252
	10000	0.06369	0.02088	0.02278	0.02485	0.01487	0.02389
100	1000	0.19670	0.02178	0.01721	0.00887	0.01573	0.01388
	10000	0.00962	0.01011	0.00449	0.00450	0.00639	0.00295
200	1000	0.01496	0.00345	0.00237	0.00248	0.00221	0.00198
	10000	0.00847	0.00139	0.00126	0.00087	0.00110	0.00127
500	1000	0.05480	0.01032	0.01423	0.00059	0.00679	0.00511
	10000	0.01608	0.00523	0.00092	0.00020	0.00028	0.00027

Table 29: Execution time of MMACS. The number of ants varies between 1 and 100.

N	R	$m1$	$m2$	$m3$	$m4$	$m5$	$m6$
50	1000	0.01501	0.03193	0.04690	0.05370	0.13695	0.23765
	10000	0.01918	0.04428	0.10274	0.11360	0.29489	0.53658
100	1000	0.06359	0.16153	0.35330	0.36749	0.60109	1.06442
	10000	0.07816	0.27944	0.50458	0.73181	1.55836	2.23375
200	1000	0.35064	0.81911	1.22310	1.64040	4.42083	8.29943
	10000	0.55900	2.20860	5.59126	4.02055	6.85303	10.5081
500	1000	5.07768	20.2446	20.4123	10.5397	34.7523	65.2062
	10000	8.64695	28.3465	44.5976	38.1680	118.269	153.611

Table 30: Number of problems solved by MMACS, MMAS, MMAS with 2opt, ACS, ACS with 2opt and QEA, in percentage

N	R	**MMACS**		**MMAS**		**MMAS-2OPT**		**ACS**		**ACS-2OPT**		**QEA**	
		BS	***PS***	***BS***	***PS***	***BS***	***PS***	***BS***	***PS***	***BS***	***PS***	***BS***	***PS***
50	1000	82.0%	52.0%	75.0%	48.0%	77.0%	47.0%	49.0%	28.0%	60.0%	33.0%	0.0%	0.0%
	10000	52.0%	13.0%	41.0%	5.0%	44.0%	5.0%	31.0%	4.0%	34.0%	5.0%	0.0%	0.0%
100	1000	90.0%	82.0%	38.0%	36.0%	45.0%	40.0%	63.0%	60.0%	69.0%	0.0%	0.0%	0.0%
	10000	63.0%	46.0%	10.0%	5.0%	14.0%	5.0%	28.0%	15.0%	22.0%	9.0%	0.0%	0.0%
200	1000	92.0%	88.0%	12.0%	12.0%	16.0%	13.0%	75.0%	73.0%	74.0%	70.0%	0.0%	0.0%
	10000	79.0%	72.0%	2.0%	2.0%	3.0%	3.0%	36.0%	32.0%	26.0%	21.0%	0.0%	0.0%
500	1000	98.0%	96.0%	2.0%	2.0%	3.0%	1.0%	65.0%	65.0%	69.0%	66.0%	0.0%	0.0%
	10000	79.0%	78.0%	0.0%	0.0%	1.0%	0.0%	29.0%	28.0%	34.0%	33.0%	0.0%	0.0%
1000	1000	**100%**	96.0%	0.0%	0.0%	2.0%	1.0%	38.0%	38.0%	40.0%	40.0%	0.0%	0.0%
	10000	90.0%	90.0%	0.0%	0.0%	2.0%	2.0%	20.0%	20.0%	20.0%	20.0%	0.0%	0.0%
2000	1000	**100%**	98.0%	0.0%	0.0%	1.0%	1.0%	40.0%	40.0%	10.0%	10.0%	0.0%	0.0%
	10000	96.0%	94.0%	0.0%	0.0%	0.0%	0.0%	30.0%	30.0%	0.0%	0.0%	0.0%	0.0%

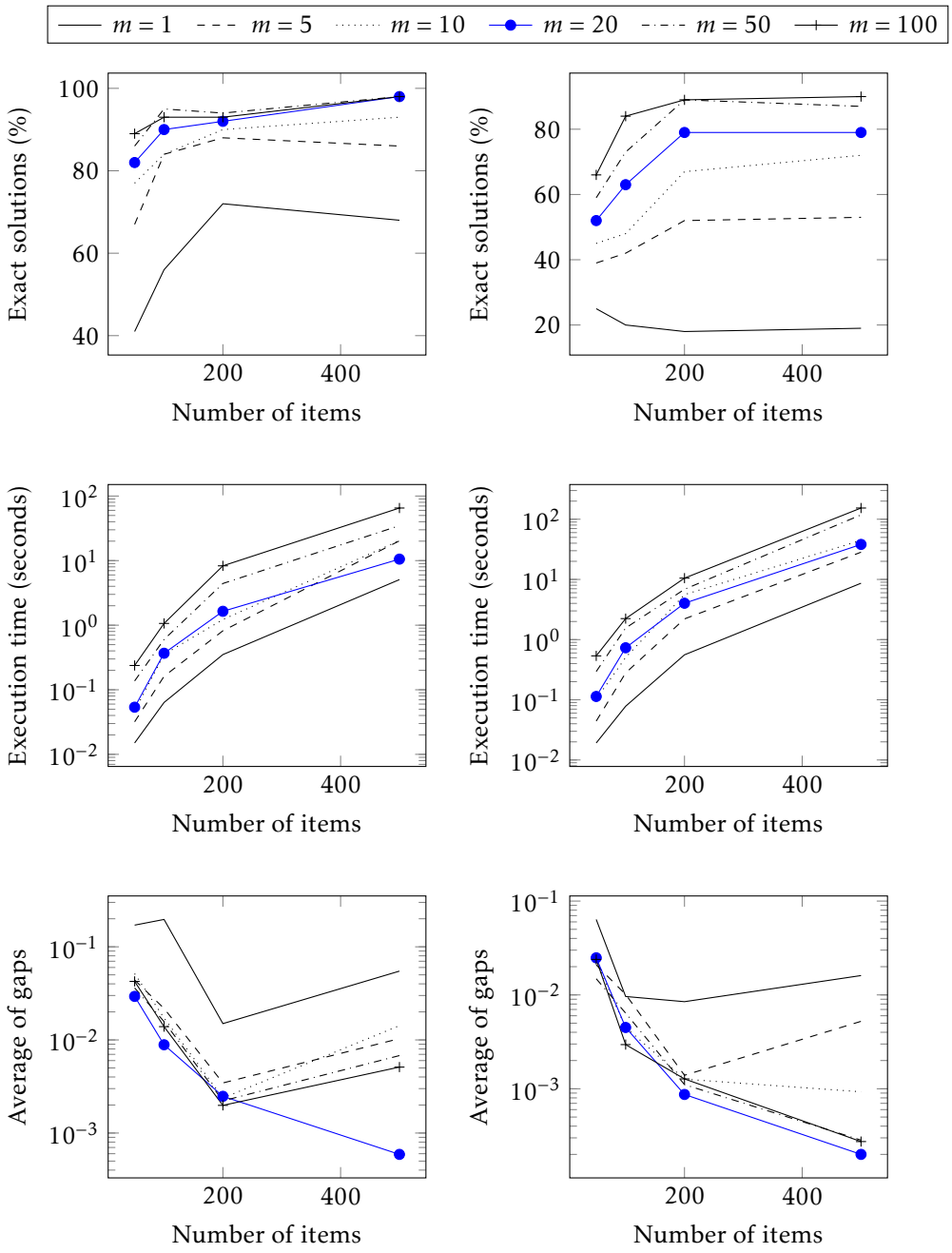


Figure 5: MMACS with various values of the number of ants. Plots on the left show results for SCKP instances with a range of coefficients equal to 1000 and plots on the right show results for SCKP instances with a range of coefficients equal to 10000.

Table 31: Average of GAPS of MMACS, MMAS, MMAS with 2opt, ACS, ACS with 2opt and QEA

N	*R*	**MMACS**	**MMAS**	**MMAS-2OPT**	**ACS**	**ACS-2OPT**	**QEA**
50	1000	0.02934	0.23080	0.02355	0.35243	0.02780	15.9902
	10000	0.02485	0.09708	0.02444	0.29466	0.03918	19.0041
100	1000	0.00887	0.76940	0.08039	0.13530	0.01213	39.4817
	10000	0.00450	0.50409	0.04409	0.08419	0.00915	39.6268
200	1000	0.00248	0.16937	0.16207	0.20180	0.02418	55.7324
	10000	0.00087	0.15295	0.14329	0.03192	0.00513	50.7264
500	1000	0.00059	0.33796	0.33374	0.04334	0.04135	92.6842
	10000	0.00020	0.30745	0.30680	0.02888	0.02113	91.554
1000	1000	**0.00000**	0.79903	0.42312	0.04723	0.09261	109.042
	10000	0.00015	0.83977	0.43098	0.10385	0.15489	115.452
2000	1000	**0.00000**	1.27071	0.49220	0.10819	0.09249	123.848
	10000	0.00001	1.33153	1.37336	0.08955	0.16696	130.323

Table 32: Execution time of MMACS, MMAS, MMAS with 2opt, ACS, ACS with 2opt and QEA, in seconds

N	*R*	**MMACS**	**MMAS**	**MMAS-2OPT**	**ACS**	**ACS-2OPT**	**QEA**
50	1000	0.05370	0.10212	0.10072	0.11601	0.11575	0.73706
	10000	0.11360	0.25768	0.15341	0.15797	0.14653	0.70629
100	1000	0.36749	1.72373	1.06384	0.68345	0.55707	2.30542
	10000	0.73181	1.37522	1.38279	1.10934	1.06852	2.31311
200	1000	1.64040	19.1191	9.92326	3.65467	3.44395	8.40170
	10000	4.02055	11.0615	10.2807	7.45693	8.76620	8.21929
500	1000	10.5397	189.923	182.438	67.8814	61.3986	46.9157
	10000	38.1618	195.993	164.348	131.969	127.623	46.3203
1000	1000	47.9736	829.651	1284.21	898.149	326.323	174.192
	10000	191.046	867.371	1281.49	730.028	420.990	183.470
2000	1000	340.109	3332.80	1437.79	3648.84	4474.66	722.961
	10000	1318.38	3359.86	3152.59	4253.97	4566.97	781.133

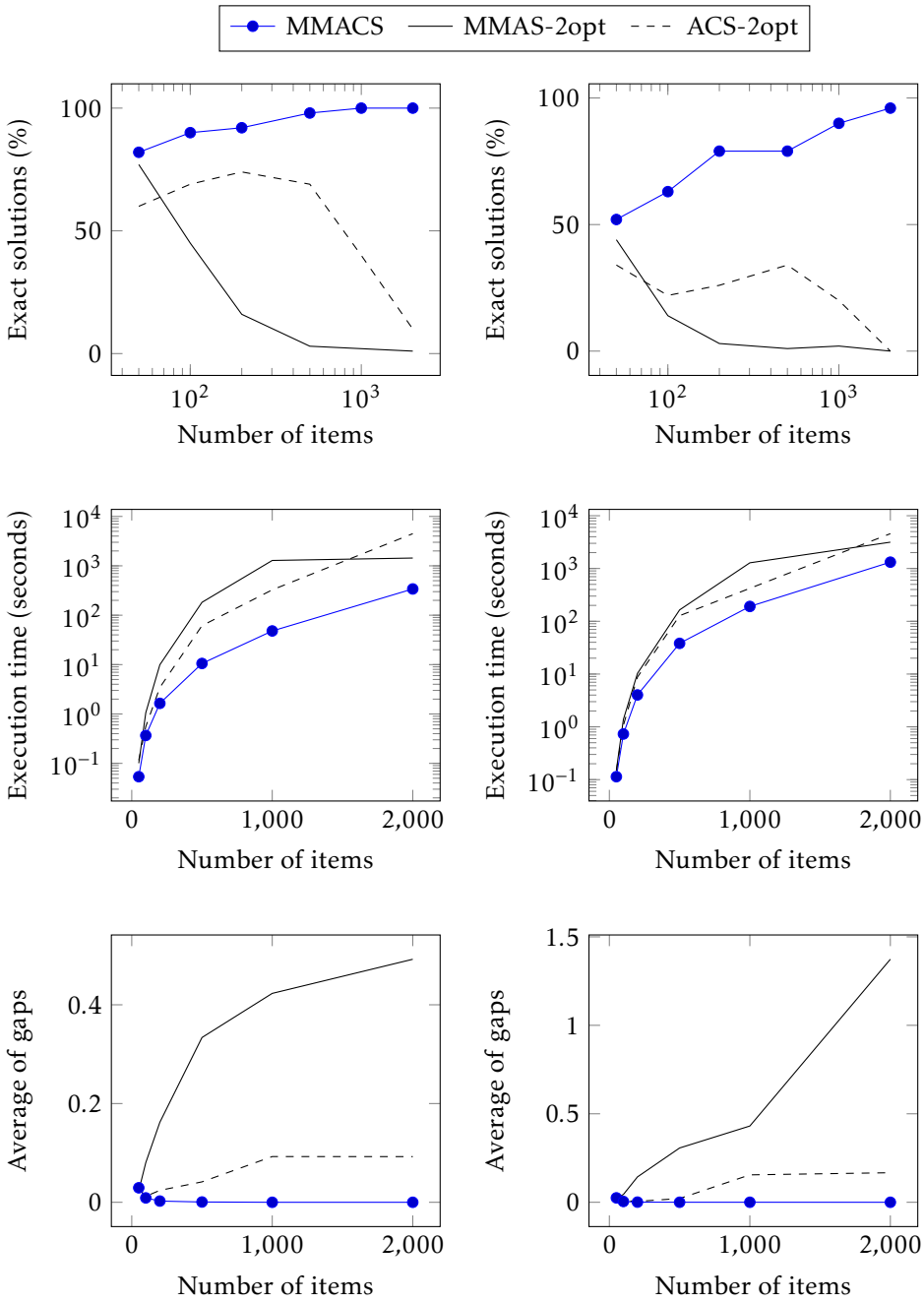


Figure 6: Comparison of MMACS, MMAS with 2-opt and ACS with 2-opt results. Plots on the left show results for SCKP instances with a range of coefficients equal to 1000 and plots on the right show results for SCKP instances with a range of coefficients equal to 10000.

reduce the time. The values 20, 50 and 100 give almost the same percentages. Thus, the value 20 gives the best compromise between a reasonable execution time and acceptable solutions. The solutions are satisfying regarding the considerable percentage of exact solutions and the reduced gap values.

6.3 MMACS experimental settings

We have fixed the parameter values after a set of experimental tests. We have set α to 1, β to 5 and ρ to 0.02, where α and β are the two parameters that determine the relative importance of the pheromone and the heuristic factors and ρ is the evaporation factor. The number of cycles were set to 20 and 20 is the number of ants. As for pheromone trails bounds, we have set τ_{max} to 6 and τ_{min} to 0.01. Finally, the fixed parameter q_0 was set to 0.9.

6.4 MMACS results

After fixing the parameter values of MMACS, empirical results are presented in this section. At a first stage, MMACS results are compared with those of MMAS [9, 10] and ACS [11]. After that, the performances of MMACS while solving SCKP are evaluated and compared to recent methods in the literature: the 2-optimal heuristic [5] and the QEA algorithm [6].

6.4.1 Comparison of MMACS, MMAS and ACS

We test MMACS and the two ACO algorithms: MMAS [9, 10] and ACS [11]. Then, we compare the obtained results. Table 1 gives the default values of the ACO parameters. MMACS, MMAS and ACS were performed on Pisinger instances of size 50, 100, 200, 500, 1000 and 2000. MMAS and ACS were tested with and without the employment of a local search. Results are given in Tables 30- 32. Their performance was compared in terms of percentage of exact and perfect solutions in Table 30, deviation of the best solution found from the optimal solution in Table 31 and the execution time in Table 32. Then, the three ACO algorithms are compared in Figure 6. Results show that for all instances, MMACS algorithm outperform the two ACO algorithms in terms of percentage of exact solutions, execution time and deviation of the best solution found from the optimal solution.

6.4.2 Comparison of MMACS with state of art algorithms

Experiments were conducted on two sets of instances and presented in this section. In the first part of experiments, MMACS, 2-optimal heuristic and QEA solve the instances of the Pisinger set. In the second part, MMACS an QEA were used to solve the instances of the generated set.

Experimental results on the instances of the Pisinger set: We present in this part the results of the experiments realized on the Pisinger set instances of the Strongly Correlated Knapsack Problems. Table 33 shows that in most cases, MMACS turned out to outperform both state of art algorithms. In fact, QEA could not solve these problems to optimality, unlike 2-optimal heuristic which showed better results than MMACS in one case out of four. Besides, our proposed algorithm MMACS reached one hundred percent of solved problem starting with the number of items equal to 1000 and a range of coefficients equal to 1000.

Table 33: Percentage of problems solved by 2–optimal heuristic, MMACS algorithm and QEA algorithm

N	R	2-Optimal	QEA	MMACS
100	1000	68.9%	0%	**90%**
	10000	13.9%	0%	**63%**
1000	1000	99.6%	0%	**100%**
	10000	**96.7%**	0%	90%
2000	1000	**100%**	0%	**100%**
	10000	**100%**	0%	**100%**

Experimental results on the instances of the generated set: Additional experiments were conducted on the instances of the generated set. We compare the results of the MMACS algorithm with a state of art algorithm QEA [6]. In QEA, the population size, the maximum number of generations, the global migration period in generation, the local group size and the rotation angle were set to 10, 1000, 100, 2 and 0.01 π, respectively.
Both algorithms MMACS and QEA were run under the same computational conditions on instances of the generated set.
In experiments of this part, the SCKP numeric parameters were set to the following values: $R = 10$, $k = 5$ and the number of items are 100, 250 and 500. The generated instances used here are similar to those presented in [6].
The exact solutions of generated instances were obtained using a dynamic programming algorithm [16] that we implemented.
Table 34 shows that MMACS found 30/30 exact solutions for all instances where MMACS BFS (best found solutions) are equal to the optima. Those significant results were given within an acceptable execution time when compared with QEA. The execution time (CPU) and the gap between the found solution and the optimum (Gap) were averaged over 30 runs.
Besides, MMACS and QEA were compared using Wilcoxon Signed Rank Test [17]. This nonparametric test shows that the two groups of data are different according to z-statistic value and p-value at the 0.01 significance level, as shown in Table 35.

Table 34: Comparative results of MMACS algorithm with a state of art algorithm QEA

N	**QEA**				**MMACS**			
	BFS	***Gap***	***CPU***	***BS***	***BFS***	***Gap***	***CPU***	***BS***
100	572	7.43099	1.04186	0%	**607**	**0.00**	**0.00423**	**100%**
250	1407	11.7935	4.48270	0%	**1547**	**0.00**	**0.05806**	**100%**
500	2115	20.5593	15.5481	0%	**2499**	**0.00**	**0.98446**	**100%**

Table 35: Comparative results of MMACS and QEA using Wilcoxon Signed Rank Test

N	**z-satistic**	**p-value**
100	-4.7821	0.00000
250	-4.7821	0.00000
500	-4.7821	0.00000

7 Conclusion

The paper presents a comparative study of the proposed hybrid algorithm MMACS while solving the strongly correlated knapsack problems. We gave an experimental analysis of the impact of different parameters on the behavior of MMACS algorithm. Experiments show that the default parameter settings proposed in the literature, gave the best possible results essentially in terms of execution time. It is also noticed from the results that ants in MMACS construct solutions relying mainly on the heuristic information rather than pheromone trails. In fact, initializing pheromone trails to the upper bound helps ants to start the search in promising zones. Besides, the MMACS balances between exploitation and exploration by the employment of a choice rule that alternate between greedy and stochastic approaches. Then, MMACS results were compared to those of MMAS and ACS, the three algorithms show very different behaviors when solving SCKP. The MMACS algorithm outperforms both ant algorithms. In the second part of experiments, we compared MMACS to other recent metaheuristics. Basically, MMACS gave better quality of solutions. As perspective, we propose to draw more attention to the exploitation of the best solutions, in order to avoid early search stagnation. This can achieve the best performances of MMACS.

Conflict of Interest The authors declare no conflict of interest.

References

[1] W. Zouari, I. Alaya, M. Tagina, "A hybrid ant colony algorithm with a local search for the strongly correlated knapsack problem" in 2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), Hammamet Tunisia, 2017. https://doi.org/10.1109/aiccsa.2017.61

[2] D. Pisinger, "Core problems in knapsack algorithms" Oper. Res., 47 (4), 570–575, 1999. https://doi.org/10.1287/opre.47.4.570

[3] T. Cormen, C. Leiserson, R. Rivest, C. Stein, "Greedy algorithms" Introduction to algorithms, 1990. https://doi.org/10.2307/2583667

[4] Stützle, Thomas and Ruiz, Rubén, Iterated greedy, Handbook of Heuristics, Springer, 2018. *https* : *//doi.org*/10.1007/978 – 3 – 319 – 07124 – $4_1$0

[5] D. Pisinger, "A fast algorithm for strongly correlated knapsack problems" Discrete Appl. Math., 89 (1-3), 197–212, 1998. https://doi.org/10.1016/s0166-218x(98)00127-9

[6] K.-H. Han, J.-H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization" IEEE Trans. Evol. Comput., 6 (6), 580–593, 2002. https://doi.org/10.1109/tevc.2002.804320

[7] K. O. Jones, "Ant colony optimization, by marco dorgio and thomas stützle" ROBOTICA, 2005. https://doi.org/10.1017/s0269888905220386

[8] Dorigo, Marco and Stützle, Thomas, Ant colony optimization: overview and recent advances, Handbook of metaheuristics, Springer, 2019. *https* : *//doi.org*/10.1007/978 – 3 – 319 – 91086 – $4_1$0

[9] T. Stützle, H. Hoos, "Max-min ant system and local search for the traveling salesman problem" in IEEE International Conference on Evolutionary Computation, Indianapolis USA, 1997. https://doi.org/10.1109/icec.1997.592327

[10] T. Stützle, H. H. Hoos, "Max–min ant system" Future Gener. Comput. Syst., 16 (8) 889–914, 2000. https://doi.org/10.1016/s0167-739x(00)00043-1

[11] M. Dorigo, L. M. Gambardella, "Ant colony system: a cooperative learning approach to the traveling salesman problem" IEEE Trans. Evol. Comput., 1 (1), 53–66, 1997. https://doi.org/10.1109/4235.585892

[12] I. Alaya, C. Solnon, K. Ghedira, "Ant colony optimization for multi-objective optimization problems" in 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007), Patras Greece, 2007. https://doi.org/10.1109/ictai.2007.108

[13] G. A. Croes, "A method for solving traveling-salesman problems" Oper. Res., 6 (6), 791–812, 1958. https://doi.org/10.1287/opre.6.6.791

[14] T. Stützle, M. López-Ibánez, P. Pellegrini, M. Maur, M. M. De Oca, M. Birattari, M. Dorigo, Parameter adaptation in ant colony optimization, Autonomous search, Springer, 2011. https://doi.org/10.1007/978-3-642-21434-9_8

[15] R. W. Morrison, K. A. De Jong, "Measurement of population diversity" in International Conference on Artificial Evolution (Evolution Artificielle), Berlin Heidelberg, 2001. https://doi.org/10.1007/3-540-46033-0_3

[16] P. Toth, "Dynamic programming algorithms for the zero-one knapsack problem" Computing, 25 (1), 29–45, 1980. https://doi.org/10.1007/bf02243880

[17] F. Wilcoxon, S. Katti, R. A. Wilcox, Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test, Selected tables in mathematical statistics 1, 1970.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 1-5 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Evaluation of Methods for Sentence Similarity for Use in Intelligent Tutoring System

Emil Brajković*, Daniel Vasić, Tomislav Volarić

Faculty of Science and Education, Department of Informatics, University of Mostar, 88 000, Bosnia and Herzegovina

ARTICLEINFO

Article history:
Received: 04 July, 2018
Accepted: 20 August, 2018
Online: 05 September, 2018

Keywords:
Intelligent Tutoring Systems
Natural Language Processing
Natural Language Generation
Sentence Similarity
Vector Space Models
Tree Edit Distance
Fuzzy Analytic Hierarchy Process

ABSTRACT

Finding similarity of textual data is very important task in natural language processing. In this article we present approach to finding similarity of words, paragraphs, sentences and documents. Semantic similarity is one of the central tasks in many applications, including text summarization, Intelligent Tutoring Systems (ITS) etc. In ITS sentence similarity is used to compare the student's response with the correct answer. The result is used to gain information about student's level of knowledge. We propose three different methods that measure text to text semantic relatedness. There are multiple approaches to finding the right measure to determine the similarity of the sentences. Some measure the alignment of characters, and other measure semantic similarity between sentences. In this work we present and evaluate methods for finding not just similarity of sentences but even also similarity of whole paragraphs and documents. We have evaluated these methods using the data from the Yahoo Question and Answer of the Non-Factual Data Set.

1 Introduction

In this paper we extend our research from[1] where we presented and evaluated methods for finding similarity between textual data. These methods can be used inside modules of ITS to evaluate student performance [2]. ITS systems are computer-based instructional systems that are composed of four principal models: the expert model, the learner model, the tutor model and the interface. Expert model otherwise known as domain model is intended as a model that contains all concepts, rules and problem solving strategies. Expert model can be used in variety of contexts such as a source of expert knowledge, student's evaluation and student performance error detection. In this model components for sentence similarity are very useful. Student model is an overlay on expert model and is one of the main components of ITS system. In this model special attention is directed to student's affective and cognitive state and it's evaluation through the learning process [3]. Tutoring model receives the information from student and domain model. Tutoring model is an ITS component that makes choices about tutoring actions and strategies. One of the most important component of ITS system is communication module [4]. Main objective of this module is to fill in the gap between the computer who understands machine language instructions and user of an ITS system who understands natural language. Often in communication module question answering systems are implemented. The student receives the question about domain knowledge and answers the question in natural language. It is very important to find robust method of evaluating the students answer. Students answer to given question can be full sentence, phrase or even one word, so this task is not trivial. In general sentence similarity methods can be divided into two categories true understanding and text to text methods [5]. Existing methods for finding sentence similarity work with very high dimensional data and are not adaptable for use in ITS systems. This paper focuses directly on computing the similarity between very short texts. The focus is on finding true understanding text similarity for use in student answer evaluation. The structure of this article includes related work section followed by Proposed algorithm for sentence similarity and Evaluation of proposed methods section. In Related work section we list all similar works in this field. We also show how our proposed method differs from other

*Emil Brajković, Matice hrvatske b.b., +38763181014 & emil.brajkovic@fpmoz.sum.ba

similar methods. The section Proposed algorithm for sentence similarity includes couple subsections where we discuss the implementation of algorithm evaluated in this article. In the section Evaluation of proposed methods, we expand our previous evaluation with the new method explained in this article. In this section we briefly comment our evaluation methodology and show our results. In the conclusion we list our conclusions and future work related to this article.

2 Related work

ITS systems that are used in many different fields[3] could greatly benefit from robust sentence similarity method. Recent advancements in natural language processing has contributed many valuable solutions in finding sentence similarity. This section reviews some of the solutions addresses the difference between them and methods proposed in this paper. Overall methods for finding sentence similarity could be knowledge based and statistical based. Some of the most popular statistical based sentence similarity methods is Latent Semantic Analysis (LSA)[6]. LSA works by applying Singular Value Decomposition to co-occurrence matrix that is generated from textual data. Resulting dense representations of words on which cosine distance can be used to determine its relatedness. LSA method can distinguish between highly similar and dissimilar words but it is insensitive to minor similarities or dissimilarities between word pairs[7]. More recent knowledge based methods for sentence similarity measurement rely on semantic databases to find distances between concepts from two compared sentences[8]. These methods relying on the existence of concepts from sentence inside the semantic database, word sense disambiguation tool and the algorithm for finding similarity of two concepts. Another approach finding sentence similarity are the methods that are based on word embeddings[9]. These methods are based on finding dense word representations that are extracted from hidden layers of neural network. The cosine distance can be used to calculate word similarity (word that appear in similar contexts are closer). This distance is very good approximation of similarity because Word2Vec might find different sorts of most-similar words. This makes Word2Vec model very scalable and usable in real world applications. Word embedding methods are more preferred over approaches using lexical databases and show better results on word similarity evaluation tasks. The state-of-the-art systems implement hybrid approach using word embeddings and lexical databases[10]. In this work we use hybrid approach for finding sentence similarity using word embeddings and non-projective tree matching algorithm.

3 Proposed algorithm for sentence similarity

In this section we provide overview of our algorithm for sentence similarity that is used for evaluation of students answers in ITS system. There are several natural language processing steps that include syntax and semantic text preprocessing. Such preprocessed text is used as a structure for our algorithm to find the list of phrases that are semantically similar. We preprocess the sentence based on it's semantic constituents. Then we build semantic tree of phrases for the two sentences that are being compared. Global steps for this approach in sentence similarity finding are:

- Text preprocessing for two sentences for which we want to find the similarity
- Finding word similarity based on word embeddings
- Finding maximum common subgraph based on association graph

The details of the implementation will be explained in subsections.

3.1 Text preprocessing

Nodes are lists of words that can be connected to other words. Globally we can divide nodes into noun nodes, verb nodes, proposition nodes and adverbial node. Nodes with multiple words contain main word which is called head word. Visual representations of the nodes in sentences are shown in fig. 01.

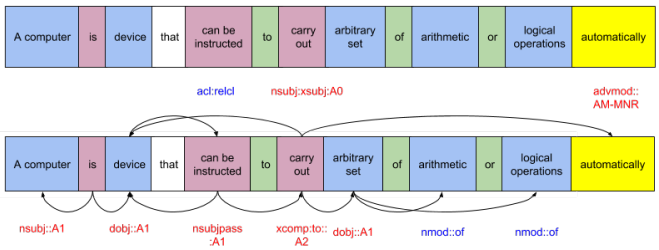


Figure 1: Parsed sentence using POS tagging algorithm for English language, node dependency relations and semantic roles

The words in sentence are segmented into nodes. These nodes are connected by syntactic and semantic relations that are parsed using CoreNLP[11] and SENNA[12] tools.

3.2 Finding similarity of words in sentences using word embeddings

To solve the word order problem the idea is to develop algorithms that will find similarity of two graphs. As shown in figures above the sentences are actually non projective trees, there are many algorithms that deal with tree similarity[13][14]. To find sentence similarity we need to compare the similarity of nodes in dependency tree of the two sentences. To find the word similarity, GloVe[15] word embeddings are used. The GloVe is actually a dictionary with word vector pairing

such as Word2Vec[16]. This dictionary is developed from training a shallow neural network on large text corpora using co occurrence matrix generated from all words in corpora, the network is optimized on custom loss function and 100 dimensional weights that are at start randomly initialized after the optimization with SGD algorithm are paired with corresponding words. Such vector representations are called word embeddings and are very popular because the structure and the way the vectors are obtained give great results for finding semantically similar words. The similarity of words can be calculated with cosine distance of two word vectors. Every word is paired with 100 dimension real valued vector that is extracted from shallow layer of neural network. Nodes that are most similar are paired to so called node pairs. Nodes are paired with nodes with maximal similarity and similarity is calculated with algorithm 01.

Data: corresponding sentences G_1 and G_2
Result: list of edges E and similarities S
while v_1 *in* G_1 **do**
  while v_2 *in* G_2 **do**
    if $similarity(v_1, v_2) > border$ **then**
      put v_1, v_2 into list of edges E;
      put sim into list of similarities S;
    else
      continue the iteration;
    end
  end
end

Algorithm 1: Algorithm for pairing nodes inside two sentences

Similarity function is something that can be greatly improved, it's very dependent on the dictionary and corpora that is trained on, the newer methods for sentence similarity include finding similarities of word phrases using approaches such as skip-thoughts[17]. For every node word sense disambiguation is applied in sentence preprocessing step so the concept based similarities can also be used, but our experiments showed better results for word embeddings approach mainly because not all phrases are contained in WordNet database[18].

Glove embeddings showed very good results for finding phrase similarity. Phrase similarity is similarity of multiple words, or in this case nodes. For calculation of phrase similarity we used Word Mover's Distance[19] Finding similar node pairs is process where borders of similarity are set to exclude the nodes that have very small similarity or no similarity at all. Borders are defined based on the word type, our assumption is that if nodes are verbs or nouns the similarity should be larger, but if the node is apposition, proposition, pronoun or something else the border of similarity should be not that high. The borders are defined as follows: Verb nodes - 0.80, Noun nodes and all other node types (pronoun, aposition, proposition) - 0.60. The similar and paired nodepairs are shown in fig. 02.

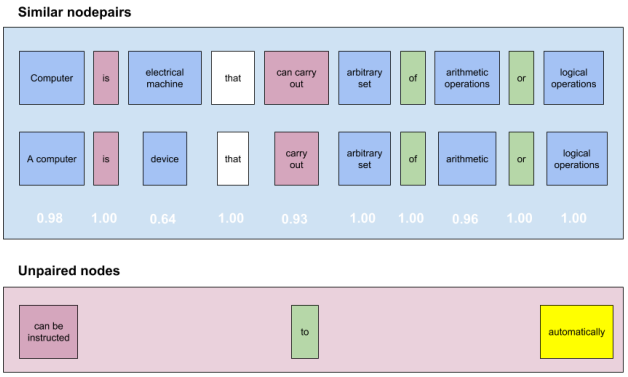


Figure 2: Paired node pairs from two sentences

3.3 Finding maximum common subgraph of whole sentences based on association graph

The next step is to find the maximum complete subgraph of two graphs, to do that we use nodes that are similar. From constructed node pairs we create so called association graph as defined in [20], all nodes that are similar are used to generate graph on which algorithm for finding maximum clique is applied.

Association graph AG The graph product $GP(V;E)$ of two graphs $G1(V1;E1)$ and $G2(V2;E2)$ is a new graph defined on the vertex set $V = V1 \otimes V2$ and the set of edges $E = E_1 \otimes E_2$. The association graph $AG(E,V)$ defined here is one of the graph products with the following adjacency conditions. Any $e = v_i v_s; v_j v_t$ considered to be adjacent:

- if $v_i \in V_{1is}$ adjacent to $v_j \in V_1$ in the original graph G_1 and $v_s \in V_2$ is adjacent to $v_t \in V_2$ in the original graph G_2 , or,
- if v_i is not adjacent to v_j and vs is not adjacent to v_t.

The association graph AG made by the previous definition should have all possibilities of vertex matches between two initial graphs $G1$ and $G2$; namely, a clique in AG corresponds to a common subgraph between $G1$ and $G2$; . Thus, the original problem of obtaining the largest match of sentences can be reduced to the computational task of searching for the largest clique in $AG, MCL(AG)$. For given two sentences the association graph is used for finding maximum clique.

A subgraph $subg_1$ of graph $G1$ is a new graph obtained from $G1$ by deleting some edges and vertices's, the same applies to $subg_2$.The nodes and vertices's that are deleted are those that do not belong to the maximum clique $MCL(AG)$.

If $(node_1, node_2) \notin AG$, where it is $node_1 \in G_1$ and $node_2 \in G_2$, then $node_1$ and $node_2$ are deleted from AG.

Subgraph $subg_1$ and subgraph $subg_2$ are maximal common subgraph MCS. For computing similarity we experimented with use of the Jaccard coefficient[23]

which can be calculated shown in equation (1).

$$JC = (G1, G2) = M\ddot{C}S(G1 + G2) - MCS \quad (1)$$

This measure is the measure for the similarity of the sentences, so if we want to get the dissimilarity measure DM we use the JC and subtract it as follows $DM = 1 - JC$. We used this measure for evaluation and comparison with other algorithms.

Data: list of all edges E and sentences G_1 and G_2
Result: list of connected edges AG
while E_1 *in* E **do**
  while E_2 *in* E **do**
    if v_1 *of* E_1 **and** v_1 *of* E_2 *are adjacent in* G_1 **then**
      if v_2 *of* E_1 **and** v_2 *of* E_2 *are adjacent in* G_2 **then**
        put E_1, E_1 into list;
      else
        continue the iteration;
      end
    else
      if v_2 *of* E_1 **and** v_2 *of* E_2 *are not adjacent in* G_2 **then**
        put E_1, E_1 into list;
      else
        continue the iteration;
      end
      continue the iteration;
    end
  end
end

Algorithm 2: Algorithm for finding association graph AG

4 Evaluation of proposed methods

Yahoo Non Factoid Question Dataset[24] was used in evaluation of good vs bad answers. The dataset contains 87,361 questions and appropriate answers, from which one of them was marked as the best answer for that question. Each question contains it's best answer along with other answers submitted by users. This dataset contains non-factoid QA, which means it covers topics beyond factoid question answering, such as "Who is the father of artificial intelligence?".

These answers are somewhat complex and finding function for similarity is a complex task. The objective was to find the method that maximizes the function of dissimilarity of two sentences. Reason for this is that answers that are labeled as best must be different than the other answers, labeled as bad answers. The evaluation methodology was conducted in three phases using three algorithms. We evaluated parse tree, knowledge tree and word vector representations. In the evaluation phase we define dissimilarity function as the function where we subtract the similarity score from maximum score. The parse tree and knowledge tree was traversed using Euler's algorithm and then Sorensen-Dice[21][22] algorithm was used to determine the similarity of trees. The results of our evaluation are shown in table I. We included the results of new proposed method described in this article. The resulting number shows the average dissimilarity measure that is generated by sum of all scores and divided by number of sentences.

Table 1: Results of two best algorithm evaluation on whole nf6l dataset

Algorithm	Number of Questions	Number of sentences	Result
Knowledge tree and Dice	5000	29723	74.078
Word2Vec sentence dissimilarity	5000	29723	68.740
New algorithm sentence dissimilarity	5000	29723	66.120
Knowledge tree and Dice	10000	61017	74.847
Word2Vec sentence dissimilarity	10000	61017	68.808
New algorithm sentence dissimilarity	10000	61017	70.234
Knowledge tree and Dice	87361	638847	76.302
Word2Vec sentence dissimilarity	87361	638847	67.778
New algorithm sentence dissimilarity	87361	638847	78.453

In our previous work[1] the method that was giving reasonable results was the knowledge tree approach. In this algorithm we used the constituency tree which was traversed using Euler's path algorithm on and nodes where compared using WordNet similarity measure. The new method in measuring sentence similarity has shown even better results for the whole corpus. Approach given in this work is insensitive to word order and semantically close terms such as synonyms, hypernyms and other. In table I it is shown that new algorithm proposed in this work gave the best results in finding dissimilarity measure between good and bad answers. In future work we plan to evaluate this method by standard means of evaluation using some of the standard tasks for sentence similarity, such as STS Benchmark[27].

References

[1] D. Vasic, E. Brajkovic, "Tree and word embedding based sentence similarity for evaluation of good answers in intelligent tutoring system", In proceedings SoftCom, 2017.

[2] T. Volarić, D. Vasić, E. Brajković Adaptive Tool for Teaching Programming Using Conceptual Maps. In: Hadźikadić M., Avdaković S. (eds) Advanced Technologies, Systems, and Applications. Lecture Notes in Networks and Systems, vol 3. Springer, 2017.

[3] S.R.D. Santos, J.L.M. Amaral, J.F.M. Amaral, "Adaptive Intelligent Systems applied to two-wheeled robot and the effect of different terrains on performance", Advances in Science, Technology and Engineering Systems Journal, vol. 2, no. 1, pp. 1-5 (2017)

[4] R. Nkambou, J Bourdeau, R. Mizoguchi, Advances in Intelligent Tutoring Systems, Springer-Link, 2010.

[5] V. Rus, M. Lintean, A. C. Graesser, D. S. McNamara, Text-to-text similarity of sentences, Applied Natural Language Processing: Identification, Investigation and Resolution, 2011.

[6] C. Papadimitriou, H. Tamaki, P. Raghavan, P. Raghavan, S. Vempala, "Latent Semantic Indexing: A Probabilistic Analysis", In proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998, pp. 159–168.

[7] S. Simmons, Z. Estes, "Using Latent Semantic Analysis to Estimate Similarity", In proceedings The 28th Annual Conference of the Cognitive Science Society, 2006.

[8] A. Pawar, V. Mago, "Calculating the similarity between words and sentences using a lexical database and corpus statistics", IEEE transactions on knowledge and data engineering, 2018.

[9] M. Tomas, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", 2013.

[10] R. Speer, J. Chin, C. Havasi, "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge", In proceedings AAAI Conference on Artificial Intelligence, 2017,

[11] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," in Association for Computational Linguistics (ACL) System Demonstrations, 2014, pp. 55–60.

[12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa. "Natural Language Processing (Almost) from Scratch", Journal of Machine Learning Research (JMLR), 2011.

[13] P. Bille. Asurvey on tree edit distance and related problems. In Theory of Computer Science, 2004., pp. 217–239.

[14] R. Yang, P. Kalnis, A. K. H. Tung, "Similarity Evaluation on Tree-structured Data", In proceedings of 2005 in SIGMOD Conference, 2005.

[15] J. Pennington, R. Socher, C. D. Manning, "GloVe: Global Vectors for Word Representation", Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Jan. 2013.

[17] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun S. Fidler, "Skip-Thought Vectors", CoRR, 2015.

[18] C. Fellbaum, "WordNet: An Electronic Lexical Database", Cambridge, MA: MIT Press, 1998.

[19] M. J. Kusner, Y. Sun, N. I. Kolkin, K. Q. Weinberger, "From Word Embeddings To Document Distances",In Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015.

[20] M. Hattori,Y. Okuno ,S. Goto, M. Kanehisa, "Genome informatics. International", In proceedings of Conference on Genome Informatics, 2003.

[21] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," Ecology, vol. 26, no. 3, 1945.

[22] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biol. Skr., vol. 5, 1948.

[23] P. Jaccard, "Jaccard coefficient similarity", New Phytologist vol. 11, 1912 pp. 37–50

[24] S. Harding, "nfL6: Yahoo Non-Factoid Question Dataset." [Online]. Available: https://ciir.cs.umass.edu/downloads/nfL6/. [Accessed: 19-May-2017].

[25] D. Y. Chang, "Applications of the extent analysis method on fuzzy AHP," Eur. J. Oper. Res., vol. 95, no. 3, pp. 649–655, 1996.

[26] C.-L. Hwang and K. Yoon, The Technique for Order of Preference by Similarity to Ideal Solution. Berlin, Germany: Springer-Verlag Berlin Heidelberg, 1981.

[27] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia (2017) "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation Proceedings of the 10th International Workshop on Semantic Evaluation", 2017.

[28] T. Landauer, P. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," Discourse Process., no. 25, 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No.3, 01-06 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Experimental Results and Numerical Simulation of the Target RCS using Gaussian Beam Summation Method

Corresponding Author*, GHANMI Helmi[1], KHENCHAF Ali[1], POULIGUEN Philippe[2], LEYE Papa Oussmane[1]

[1]*Lab*-STICC UMR CNRS 6285, ENSTA Bretagne, 29806, Brest, France

[2]*French General Directorate for Armament (DGA), 75509, Paris, France*

ARTICLE INFO

Article history:
Received: 04 April, 2018
Accepted: 18 April, 2018
Online: 07 May, 2018

Keywords :
Radar Cross Section(RCS)
Gaussian Beam Summation (GBS)
Gaussian Beam Launching (GBL)
Geometrical Theory of Diffraction (GTD)
Physical Optic (PO)
Method of Moment (MoM)

ABSTRACT

This paper presents a numerical and experimental study of Radar Cross Section (RCS) of radar targets using Gaussian Beam Summation (GBS) method. The purpose GBS method has several advantages over ray method, mainly on the caustic problem. To evaluate the performance of the chosen method, we started the analysis of the RCS using Gaussian Beam Summation (GBS) and Gaussian Beam Launching (GBL), the asymptotic models Physical Optic (PO), Geometrical Theory of Diffraction (GTD) and the rigorous Method of Moment (MoM). Then, we showed the experimental validation of the numerical results using experimental measurements which have been executed in the anechoic chamber of Lab-STICC at ENSTA Bretagne. The numerical and experimental results of the RCS are studied and given as a function of various parameters: polarization type, target size, Gaussian beams number and Gaussian beams width.

1. Introduction

In the radar frequency domain, both asymptotic and rigorous methods have been developed to model the variations of the RCS of canonical and complex targets. The rigorous methods such as Method of Moment (MoM) are based on an integral formulation, and they are served to validate the new asymptotic approaches. The asymptotic methods as Physical Optic (PO) and Geometrical Theory of Diffraction (GTD) reduce the operation number of solving of high-frequency equations as for large objects [1-3]. The asymptotic methods using the hypothesis of locally plane wave and high-frequency approximation are based on the principle of rays. The application of these methods in a complex propagation scenario is often limited by the transaction between highlighted and shadowed region and the caustic problem (except the PO method). To overcome this problem, we will apply an asymptotic technique based on Gaussian beams and we will study the RCS variation of different radar targets. The used Gaussian method named Gaussian Beam summation (GBS) has been the subject of research for several years. In fact, the solutions of Maxwell's equations and Helmholtz's wave equation as single Gaussian beams were developed in the sixties. Afterward, Babich and Pankratova have proposed a mathematical study of the integral Gaussian beams where they describe them as a representation of a scalar wave field [4-9]. This integral has been used for a mathematical study of the Green's function discontinuities in the mixed problem for the wave equation. The Gaussian Beam summation as an asymptotic approach for computing high-frequency wave fields has been developed by V. Cerveny [7] and M.M. Popov [8]. The summation of Gaussian beams allows solving some critical points of the asymptotic ray methods such as the problems related to the evaluation of wave field in singular areas.

The main goal of this work is to simulate and analyze the RCS variations of canonical and complex targets using GBS method and validate the numerical simulation results by experimental measurements. Therefore, this paper is organized as follows: Section 2 shows the physical principle and the mathematical formulation of the GBS and GBL methods. Section 3, illustrates the numerical and experimental results of RSC of different radar targets. The final section presents conclusions and future research.

*GHANMI Helmi, ENSTA Bretagne 29806 Brest, France, +33 (0) 2 98 34 87 08 & helmi.ghanmi@ensta-bretagne.org

2. Formulation and analysis of Gaussian Beam Methods

2.1. Formulation of GBS method

V. Cerveny and M.M. Popov [7-9] have developed a new technique for calculation of wave fields in high-frequency approximation. This technique is called Gaussian Beam Summation. In the GBS method, the total final field in any observation point outcomes from a set of rays that passed through his vicinity. According to V. Cerveny [7], [9-10] and M.M. Popov [8], [11], the general procedure of the GBS method consists of two compatible steps. Firstly, we derive a Gaussian beam propagating along the ray for each selected ray. Each Gaussian beam has its own contribution to the receiver. In the final step, we sum all contributions over all rays [7-9].

Before showing the basic formulation of the GBS method, we must describe the assumptions that were used to establish this formulation. We've started by considering a homogeneous and isotropic medium with an electromagnetic wave propagating (with a propagation velocity v) in this medium which is being excited by a point source. Then, we've supposed that some wave's process is described by the Helmholtz's wave equation and the point source is positioned in the origin. After that, we've solved the Helmholtz's equation in the neighborhood of rays.

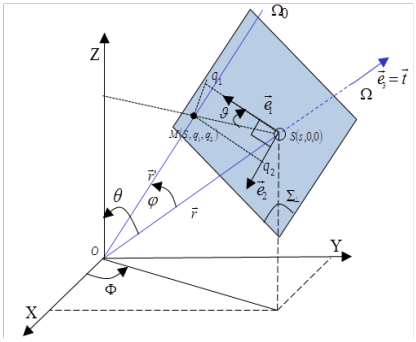


Figure 1: Geometric configuration and coordinate parameters: a point M situated in the plane $\Sigma_\perp$ perpendicular and crossing Ω at point S. The point M is located in the vicinity of ray Ω and it ray-centered coordinates are q_1, q_2, and s. The center point O is at $s=s_0$.

Figure 1 presents the ray-centered coordinate system (s, q_1, q_2) used to formulate the equation (1) of the Gaussian beam amplitude $u(s, q_1, q_2, t)$. This coordinate system is connected to any selected ray Ω. As a function of the local coordinates and at the receiver point, the solution of the Helmholtz equation as a solitary Gaussian beam is given by (1) [8].

$$\begin{cases} u(s,q_1,q_2,t)=\left(\dfrac{v}{\det[Q]}\right)^{\frac{1}{2}}\cdot\exp\left\{-j\omega\left[t-\tau(s)-\dfrac{1}{2}\left(q^T\cdot P\cdot Q^{-1}\cdot q\right)\right]\right\} \\ \tau(s)=\displaystyle\int_s\frac{ds}{v} \end{cases} \quad (1)$$

In (1) $\tau(s)$ is the travel time from the source along the selected ray, v is the propagation velocity, q^T represents the transpose of the vector, the quantities Q and P are 2 x 2 matrix called "dynamic quantities" satisfying the system ODE (2) in variations, called "dynamic ray tracing equations" (DRT) [11], [15].In a homogeneous medium with wave speed equal to the celerity c, the DRT equations can be written as:

$$\frac{dQ}{ds}=c.P;\ \ and\ \frac{dP}{ds}=0 \quad (2)$$

The system of differential equations (2), is solved by introducing initial conditions specified at an arbitrary point ($s = s_0$) on the central ray. The initial conditions are also related to three other conditions along the whole rays [12]. These conditions are:

- Even though P and Q are not symmetrical the ($P\times Q^{-1}$) must be symmetric matrix;
- $Im(P\times Q^{-1})$ is a positive-definite matrix;
- $(det[Q]\neq 0)$;

To find the initial values for Q and P, we use Hill's [13] initial data for the Green's function.

$$Q=\frac{\omega_r.\omega_0^2}{c}.I\ ;P=\frac{j}{c}.I\ \ and\ ;s=s_0 \quad (3)$$

In (3), ω_0 is the initial half beam width at the frequency $f = \omega_r/2\pi$, I is the identity matrix (2×2). Using the initial conditions in (3), we can find the general solution of (2), and can be written as follows:

$$Q=\left(\frac{\omega_r.\omega_0^2}{c}.+j.(s-s_0)\right).I\ ;\ and\ \ P=\frac{j}{c}.I \quad (4)$$

In the case of homogeneous media, by using (4) in (1), we return to the representation of the amplitude u of the Gaussian beam in 3D:

$$u(s,q_1,q_2)=\frac{c^{\frac{3}{2}}}{\omega.\omega_0^2+j.(s-s_0)}.\exp\left(j\omega\left(\tau(s)+\frac{1}{2}\frac{(q_1^2+q_2^2)}{c.\left(-j\frac{\omega.\omega_0^2}{c}.+(s-s_0)\right)}\right)\right) \quad (5)$$

Using the geometrical configuration illustrated in Figure 1, and introducing the spherical coordination system (r, θ, φ), we can deduce the following factor (in 6) as a function of the distance (r) between the transmitter and the receiver:

$$q_1^2+q_2^2=r.\sin(\varphi);\ \ and\ \ s-s_0=r.\cos(\varphi) \quad (6)$$

Finally, to calculate the full amplitude (u^{GBS}) at the receiver we must use an integral formulation as shown in (7). This integral will be calculated on all Gaussian beams described by their characteristic angle (δ) from the source:

$$u^{GBS}=\int_\delta\Phi_\varphi.u_\varphi(s,q_1,q_2).d\delta \quad (7)$$

Where, $\Phi\varphi$ is a quantity, generally complex-valued, which remains constant along the considered ray but may differ from ray to ray. It is called complex weight function. And the function $u_\varphi(s, q_1, q_2)$ is the Gaussian beam connected with the ray.

In (7) the domain δ is centered on the central ray, it delimits the beams propagating in the vicinity of the central ray, chosen in such way that the Gaussian beam $u_\varphi(s,q_1,q_2)$ outside this domain do not contribute effectively to the wave field. δ is a cone with a vertex angle φ.

$$d\delta=\sin\varphi.d\varphi.d\vartheta,\quad \vartheta\in[0,2\pi] \quad (8)$$

For a homogeneous medium, on an observation point (M) the ray asymptotic solution of the Helmholtz equation is given by the following equation:

$$u(M) = \frac{1}{4.\pi.r}.\exp\left(j.\frac{\omega}{c}.r\right) \quad (9)$$

The GBS integral, in (7), may be evaluated asymptotically using the saddle-point method. Thus, this result must coincide with the above ray asymptotic solution in the regular region [12], [17]. Matching both asymptotic solution of (7) and (9) we can determine the complex weight function $\Phi\varphi$. Integral (7) is evaluated by numerical quadrature with regular increment $\Delta\varphi$. The equation (10) is used for the numerical computation.

$$u(M) = 2.\pi.\sum_{k=1}^{N}\Phi_{\varphi k}.u_{\varphi k}.\sin(\varphi_k).\Delta\varphi_k \quad (10)$$

After the formulation of the scattered filed using GBS method ((7) and (10)), we will analyze the influence of the main parameters of the Gaussian beam on the variation on the field amplitude. Then, we will compare the solution based on Gaussian beam with the analytical solution given by (9).

2.2. Analysis of GBS method

Figure 2 compares the amplitude of field calculated by GBS method and ray asymptotic solution of the Helmholtz equation. This simulation (Figure 2) has been realized as a function of the distance (r in km) from the source to the receiver. The frequency equal to 10GHz, the beam width (ω_0) equal to 18λ (where λ is the wavelength) and the values of beams number (N) are : {133, 200, 400 and 600}. Magenta, red, green, blue and lines correspond to the GBS solution for different beams number, respectively 133, 200, 400 and 600 over which the summation is down. The black line represents the ray asymptotic solution. We can observe, that the beam density in the vicinity of the central ray offers satisfactory accuracy. In fact, when the number of the beams is more than 200, the GBS and the ray asymptotic are in good agreement. So, as with the usual techniques of ray tracing, a high beam density (200 for this case) is necessary for high accuracy.

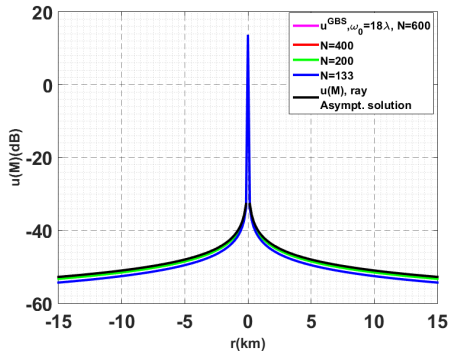


Figure 2: Comparison between the ray asymptotic solution and GBS method for beam number N={133, 200, 400, 600} and beam width is 18λ.

In Figure 3, we compare the percentage error between the ray asymptotic solution and the GBS method. We can see even at a distance of 15 km the percentage error between the ray asymptotic solution and the GBS technique is lower than 10% for a beam density N = 200 and remains below 4% fo N = 400. In addition, one should note that the computations by GBS method exhibit no singularities when passing by the source point (r=0), unlike the ray asymptotic solution. This result confirms that by using the GBS method we can overcome some limitation of the ray asymptotic models. The proof of this result relies on the theory of systems of linear first order differential equations [14-16].

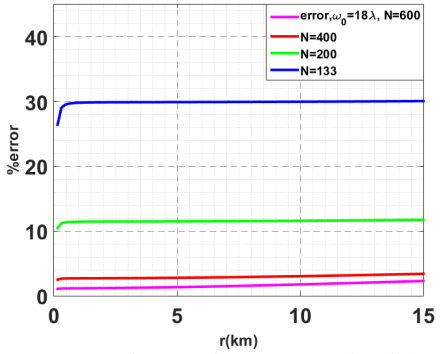


Figure 3: Percentage error between the ray asymptotic solution and the GBS approximation of the Helmholtz's wave equation for different beam number values.

Figure 4 illustrates the field amplitude calculated by GBS method (where N = 600 beams) and ray asymptotic solution. The different color lines correspond to the GBS solution for different beam width, {5λ, 8λ, 12λ, 15λ and 18λ}. This comparison shows that the GBS is ω_0 dependent. This beam initial parameter must be chosen optimally to guarantee sufficient accuracy.

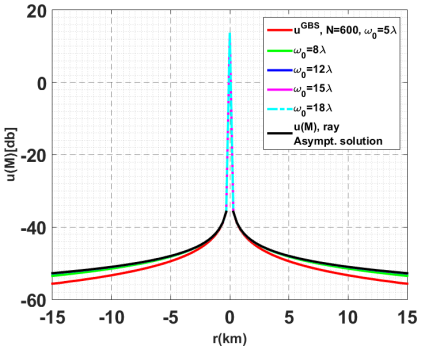


Figure 4: Comparison between the ray asymptotic solution and GBS method for different beam width values: ω_0={5, 8, 12, 15, 18} and beam number equal to 600.

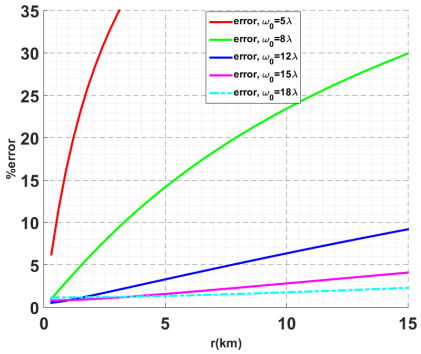


Figure 5: Percentage error between the ray asymptotic solution and the GBS approximation of the Helmholtz's wave equation for different beam width values.

To study in greater details the effect of the ω_0 parameters on the scattered field variation, we have simulated in Figure 5 the relative error between GBS solution and ray asymptotic solution. The relative error is normalized by ray asymptotic solution. The numerical analysis in Figure 5 indicates that for ω_0 = 18λ the

relative error remains below 3% even at a distance 15 km (and less than 2% at 5 km) from the source and the GBS solution match the ray asymptotic solution.

In the present paper, the GBS method was compared with another Gaussian approach called Gaussian Beam Launching (GBL). The formulations of GBL method are presented in the following part.

2.3. Formulation of GBL method

The Gaussian Beam Launching (GBL) technique has been introduced and applied in the research published by H. T. Chou et al [17]. Consider a target (plate, disc, cylinder,…) illuminated by a Gaussian beam, the GBL method is applied to calculate the radiation integral of the target scattered fields. For the considered Gaussian beam, the incident magnetic field is given by the following form [4], [17]:

$$H_i(r_i) = H_i(0)\sqrt{\frac{\rho_i + jb_i}{z_i + \rho_i + jb_i}}.\exp\left(-jk.\left(z_i + \frac{x_i^2 + y_i^2}{z_i + \rho_i + jb_i}\right)\right) \quad (11)$$

In (11), the distance between a point on the illuminated surface and the waist the incident Gaussian beam is denoted by ρ_i, the position vector in the Gaussian beam is defined by and $b_i = k.\omega_0^2/2$, where k, ω_0 are the wave number and the half beam-width respectively.

The electric fields scattered from the target surface (Σ) illuminated by the incident field is given by the integration of the incident Gaussian beam on the reflector surface (PO integral). This integral is written by (11):

$$E(r) = \frac{j.k}{4.\pi}\iint_{\Sigma}\left[\frac{Z_0}{\sqrt{\varepsilon_r}}.R \times R \times (e_z \times H_i(r_i))\right]\frac{\exp(-j.k.R)}{R}.ds \quad (12)$$

Finally, by using (11) in (12) and solving the integral, we can compute the scattered field applying GBL formulation as in [4], [17].

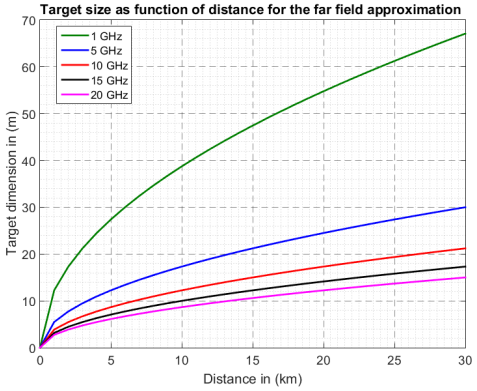


Figure 6: Target size as a function of distance for far-field approximation: The Fraunhofer criterion of far-field.

In both GBS and GBL methods, it's required to take into account the far-field approximation parameters (distance r, target size D, and wavelength λ). In order that the phases and amplitudes of the waves arriving from different regions of the target do not vary considerably with the distance (r), the far-field region must be far enough away from the source. This region of the far field starts at a distance *"r"* given by the following equation (Fraunhofer criterion) [19]: $r \geq ((2.D^2)/\lambda)$.

Figure 6 shows the minimum far-field range as a function of the target dimensions D, and for different frequencies values. In this work, the simulations and measurements are realized for two frequency bands (5GHz and 10GHz).

3. Numerical and Experimental Results

3.1. Experimental setup

The experimental measurements of RCS have been carried out on an anechoic chamber (8m x 5m x 5m) at Lab-STICC ENSTA Bretagne (see Figure 7). The characteristics of various components of measurements system are:

- All walls are covered with absorbent material;
- The transmitter and receiver antennas are identical and there polarization is horizontal or vertical. They are placed on an angular rail that allows changing their position;
- A computer controls the Vectorial Network Analyzer (Anritsu 37347D) which operates in the frequency range from 40MHz to 20GHz and the positioning system;
- The NEWPORT positioning system with an angular resolution equal to 0.01° and an angle vary between -90° and 90°. An elevation motor for adjusting the height of the target;

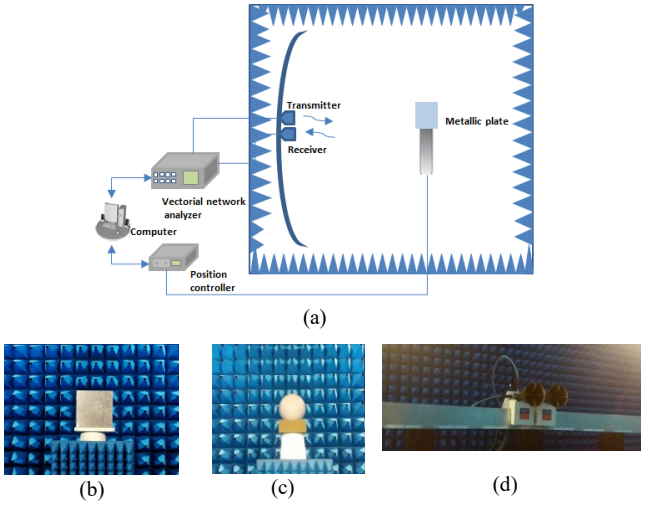


Figure 7: (a) General description of the experimental setup, (b) Target plate, (c) Radar sphere calibration, (d) Transmitter and receiver antennas.

3.2. Numerical simulation and experimental measurement of RCS

In the simulations illustrated in Figure 8, we set the azimuth angles φ_i to zero; the incident angle θ_i varies from −10° to 10° and the other acquisition parameters are: f = 10GHz, the beam width $\omega_0 = 2\lambda$, beams number N = 200. The GBS is compared with the Gaussian GBL method, the asymptotic PO model, and the rigorous MoM method (in FEKO). Comparing the curve of RCS using GBS is in the blue line, GBL technique in red line with the other models; we observe that they match rigorously the PO solution and MoM for the main beam. In fact, GBS and GBL accurately model the main beam, without mitigation, the diagrams are also consistent,

and the modeled peak width and location match the PO and MoM solutions.

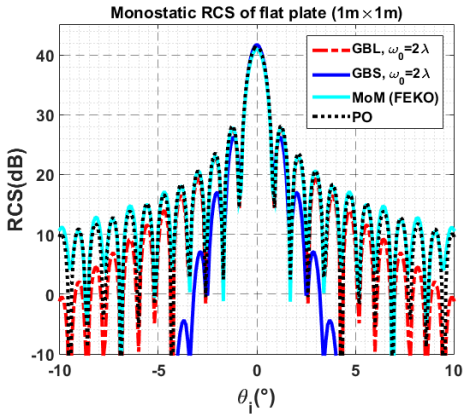


Figure 8: Monostatic RCS of a PEC flat plate computed using GBS, GBL techniques, for beam width $\omega_0 = 2\lambda$ and compared with PO and MoM methods: f =10GHz, *vv* polarization.

From Figure 8 we can conclude that the GBS method models very well the variation of the scattered field in the specular direction. However, we observe that the GBL method simulates the secondary lobes better than the GBS method. The deviation between the curves in outside specular direction can be considered through broadening the range of validity of the GBS by taking into account the edge diffraction. To consider the edge diffraction contribution, we need to use the method of the Geometric Theory of Diffraction (GTD) [18]. In fact, the diffracted field when incident field strikes the edges is calculated from the GTD and is accounted in the complex weight function in the integral (7).

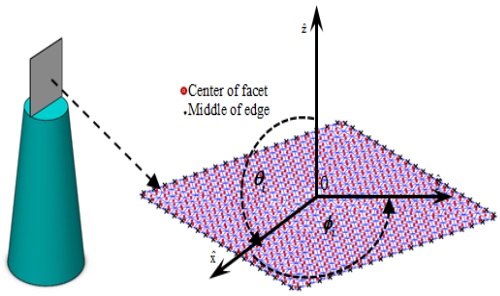


Figure 9: Geometry of a flat plate (30 cm ×30 cm) meshed with triangular patches.

Figure 9 represents the geometry of a flat plate meshed with a triangular patch (using CATIA software). In this geometric configuration, each facet is represented by a triangle node (in blue) with a central point in red color (bright points) and black cross in the middle of the external edge. This geometric configuration and this mesh will be used in the calculation of RCS by the two techniques GBL and GBS + GTD.

The simulated (using GBS, GBL, PO, and MoM) and measured RCS results of PEC plate are shown in Figure 10 and Figure 11. The acquisition parameters are: frequency of 10GHz, the incident angle θ_i varies from −60˚ to 60˚, the azimuthal scattering angle equal to 0°, the size of a flat plate as function of the wavelength is (10λ x10λ), the beam width of $\omega_0 = 2\lambda$, beam number density equal to 200 and the polarizations are *hh* and *vv* in Figure 10 and Figure 11 respectively.

The comparison between GBS, with accounting and without accounting the edge diffraction contribution, is illustrated in Figure 10. The evaluation GBS technique is also performed with GBL, PO, and MoM methods. From this comparison results, we can see that when the diffraction is accounted we obtain values of the RCS which get closer to those given by the experimental measurements and the rigorous MoM method. In addition, where the polarization state is *hh* and *vv* respectively, we remark that the PO model is insufficient in computing of edge plate diffraction contributions, although the agreement of GBS+GTD and GBL with the rigorous MoM solution and measured data are good in the most of scattering angles (particularly in *vv* polarization).

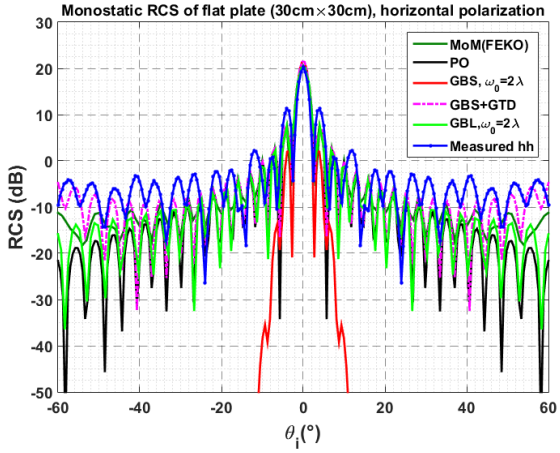


Figure 10: Comparison between GBS, GBS+GTD, GBL, the numerical models and the experimental measurements in *hh* polarization: (30cm × 30cm).

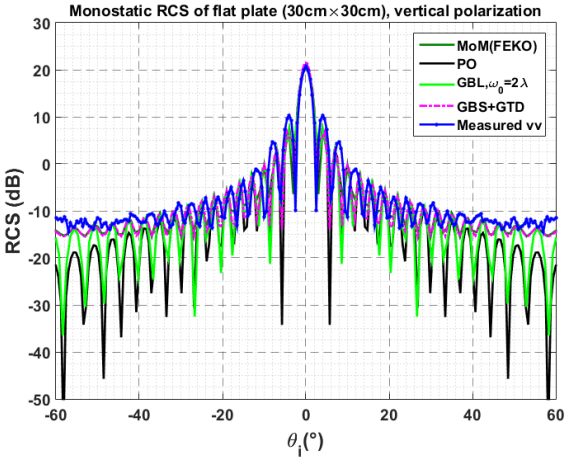


Figure 11: Comparison between GBS+GTD, GBL, the numerical models and the experimental measurements in *vv* polarization: (30cm ×30cm).

The experimental validation results show that GBS method gives a higher accurate representation of the scattered field and offer very interesting perspectives for complex targets such as cavity and corner for example. Figure 12 shows the evolution of the RCS of a metallic dihedral corner reflector realized using Ray Launching-Geometrical Optics (RL-GO) and MoM methods and also experimental measurement at a frequency of 5GHz.

In Figure 12, we can see that the curve of measured RCS is near to that simulated by MoM method. These results will serve as a basis for evaluating the development of the GBS method in our future work.

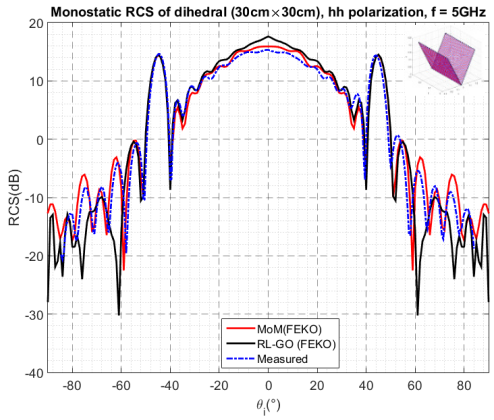


Figure 12: RCS of rectangular dihedral corner reflector (f = 5GHz): Experimental and numerical models.

4. Conclusion and future work

In this study, the RCS of radar targets has been investigated by using a new technique called Gaussian Beam Summation (GBS). In the GBS technique, the total field at the receiver is represented by the integral over all Gaussian beams propagating in the vicinity of the receiver. To study the performance of the chosen GBS method, we have carried out it theoretical formulation and we have study the influence main Gaussian parameters (beams width, the density of beams number) on the field amplitude. Then, we have introduced a numerical simulation of the RCS of a PEC plate using GBS method. The results obtained using GBS were compared with those simulated trough GBL, PO and MoM methods. In addition, we have presented the experimental measurements of RCS of canonical and complex targets. The results of RCS using GBS method were compared and validated by the experimental measurements.

The study of the RCS of different complex objects (including dihedral and trihedral corner reflector) using GBS method is one of the perspectives of the work presented in this paper.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We would like to thank the DGA (Direction Générale de l'Armement, France)-MRIS for their support of the SOFAGEMM project, where this work is in progress. Acknowledgments are also addressed to Mr. Hervé TRÉBAOL (Specialized Instructor Mechanical Design in ENSTA Bretagne) for these collaborations with us.

References

[1] F.Weinmann, "Ray tracing with PO/PTD for RCS modeling of large complex objects," IEEE Trans. Antennas and Propagation., **54**(6), 1797–1806, Jun. 2006. DOI: 10.1109/TAP.2006.875910

[2] R. Harrington, "Field computation by moment methods," Wiley-IEEE Press, 1993.

[3] H. T. Chou and P. H. Pathak, "Uniform asymptotic solution for electromagnetic reflection and diffraction of an arbitrary Gaussian beam by a smooth surface with an edge," Radio. Science., **32**(14), 1319-1336, 1997. DOI: 10.1029/97RS00713

[4] P.O. Leye, A. Khenchaf, and P. Pouliguen, "The Gaussian Beam Summation and the Gaussian Launching Methods in Scattering Problem," J. Elect. Analysis and Applications., **8**, 219-225, 2016. DOI: 〈 10.4236/jemaa.2016.810020 〉

[5] H. Ghanmi, A. Khenchaf, P. Pouliguen and P. O. Leye, "Study of RCS of complex target: Experimental measurements and Gaussian beam summation method," IEEE Conference on Antenna Measurements & Applications (CAMA), Tsukuba, pp. 196-199, 2017. DOI: 10.1109/CAMA.2017.8273399

[6] M. Katsav and E. Heyman, "Gaussian Beam Summation Representation of Beam Diffraction by an Impedance Wedge: A 3D Electromagnetic Formulation Within the Physical Optics Approximation," IEEE Trans. Antennas Propagation., **60**(12), pp. 5843-5858, 2012. DOI: 10.1109/TAP.2012.2207694

[7] V. Červený, "Summation of paraxial Gaussian beams and of paraxial ray approximations in inhomogeneous anisotropic layered structures," Seismic waves. Complex 3-D Structures., **10**, 121–159, 2000. https://doi.org/10.1111/j.1365-246X.2009.04442.x

[8] M. M. Popov, "A new method of computation of wave fields using Gaussian beams," Wave Motion., **4**, 85-97, 1982. https://doi.org/10.1016/0165-2125(82)90016-6

[9] V. Červený, "Expansion of a Plane Wave into Gaussian Beams," Stidia geoph. And geod., **26**, 120 - 131, 1982. DOI: 10.1007/BF01582305

[10] V. Červený, "Seismic Ray Theory," Cambridge: Cambridge University Press, 2001.

[11] M. M. Popov, "New method of computation of wave fields in high-frequency approximation," J. Soviet Math., **20**(1) ,1869–1882, 1982. https://doi.org/10.1007/BF01119372

[12] V. Červený, "Gaussian beam synthetic seismograms," Journal. Geoph., **58**, 44-72, 1985.

[13] N. R. Hill, "Gaussian beam migration," Geophysics, **55**(11), 1416-1428, 1990. https://doi.org/10.1190/1.1442788

[14] V. Červený, «Computation of wave field in homogeneous media,» Geophys. J. R. astr. Soc., **70**, 109-128, 1982. DOI: 10.1111/j.1365-246X.1982.tb06394.x

[15] B. S. White, A. Norris, A. Bayliss and R. Burridge, "Some remarks on the Gaussian beam summation method," Geoph. J. Royal. Astro. Society., **89**, 579-636 , 1987. https://doi.org/10.1111/j.1365-246X.1987.tb05184.x

[16] B. Bleistein, "Mathematics of Modeling, Migration and Inversion with Gaussian Beams," Colorado, USA, 2008. DOI: 10.3997/2214-4609.201405084

[17] H. T. Chou, P. Pathak and R. J. Burkholder, "Novel Gaussian Beam Method for the Rapid Analysis of Large Reflector Antennas," IEEE Trans. Antennas. Propagation., **49**(16), 880-893, 2001. DOI**:** 10.1109/8.931145

[18] J. B. Keller, "Geometrical Theory of Diffraction," J.Optical Society of America, **52**, 116 - 130, 1962. https://doi.org/10.1364/JOSA.52.000116

[19] Warren L. Stutzman, Gary A. Thiele. "Antenna Theory and Design". John Wileys & Sons, Inc., 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 2, 01-07 (2019)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

The Chimera and the Cyborg

Hybrid Compute: In vivo HPC, Cloud and Container Implementations

Lev Lafayette[*,1], Bernd Wiebelt[2], Dirk von Suchdoletz[2], Helena Rasche[3], Michael Janczyk[2], Daniel Tosello[1]

[1]*University of Melbourne, Department of Infrastructure, Melbourne, Australia*

[2]*University of Freiburg, High Performance Computing, Freiburg, Germany*

[3]*University of Freiburg, de.NBI, Freiburg, Germany*

ARTICLEINFO

Article history:
Received: 15 August, 2018
Accepted: 03 February, 2019
Online: 10 March, 2019

Keywords:
Hybrid HPC/Cloud
High Performance Computing
Cloud Computing
High Throughput Computing
Containers

ABSTRACT

High Performance Computing (HPC) systems offer excellent metrics for speed and efficiency when using bare metal hardware, a high speed interconnect, and massively parallel applications. However, this leaves out a significant portion of scientific computational tasks, namely high throughput computing tasks that can be trivially parallelized and scientific workflows that require their own well defined software environments. Cloud computing provides such management and implementation flexibility at the expense of a tolerable fraction of performance. We show two approaches to make HPC resources available in a dynamically reconfigurable hybrid HPC/Cloud architecture. Both can be achieved with few modifications to existing HPC/Cloud environments. The first approach, from the University of Melbourne, generates a consistent compute node operating system image with variation in the virtual hardware specification. The second approach, from the University of Freiburg, deploys a cloud-client on the HPC compute nodes, so the HPC hardware can run Cloud-Workloads using the scheduling and accounting facilities of the HPC system. Extensive use of these production systems provide evidence of the validity of either approach.

1 Motivation

Modern research institutions increasingly employ digital methods and workflows, which require a corresponding increase in the amount of computational resources. As data sets grow the researchers' own machines or large workstations no longer suffice, and they must turn to other resources such as High Performance Compute (HPC) and Cloud resources. Procurement, installation, and operation of these compute resources are demanding tasks which cannot be handled by individual researchers and work groups any more. Centralization of resources and administration can leverage economies of scale, but comes with compromises regarding hardware and software configurations that can cater to the needs of a growing user base. Currently neither HPC nor Cloud computing in isolation can mesh the the demands of users with the needs and resources of the compute centers.

University centers must find new, efficient ways to cater to user desires of tailored infrastructures that meets their computational needs. Any new solution should provide comparable offerings regarding features and pricing while avoiding overstretching existing personnel resources. Often demands for hardware arise with short notice and for project durations well below the normal cost amortization period of five to six years. The typical challenges of university computer centers are rooted in the very diversity of their scientific communities; wide ranging requirements for software, services, workflows, and compute resources.

Virtualization is a key technology from two fronts; it can permit accommodating diverse user requirements with a largely centralized resource, and it can help to isolate the different software environments and hardware configurations. As many resources in research infrastructures are inconsistently utilized, tapping into cloud strategies can help to significantly save on energy, personnel investment, and hardware resources.

[*]Corresponding Author: Lev Lafayette, University of Melbourne, lev.lafayette@unimelb.edu.au

https://dx.doi.org/10.25046/aj040201

This paper explores the possibility to overcome the dichotomy of cloud and HPC with a single cohesive system, and how such a setup can provide the performance of an HPC system as well as the flexibility of a cloud compute environment. Furthermore, the paper discusses if such a single system can deliver the best possible result for overall throughput and make better use of computational resources. We examine two practical approaches to make HPC resources available in a dynamically reconfigurable hybrid HPC/Cloud setup, both of which can be achieved with few modifications to existing HPC/Cloud environments. The first approach (University of Melbourne) generates a consistent compute node operating system image with variation in the virtual hardware specification. The second approach (University of Freiburg) deploys a cloud-client on the HPC compute nodes, so the HPC hardware can run Cloud-Workloads for backfilling free compute slots. This paper is an extension of work originally presented in the 2017 IEEE 13th International Conference on e-Science [1], with a new exploration of container implementations in such environments.

2 The HPC/Cloud Conflict

HPC systems running massively parallel jobs require a fairly static software environment, running on bare metal hardware, with a high speed interconnect in order to reach their full potential. Even then, they only offer linear performance scaling for cleverly designed applications. Massively parallel workloads need to be synchronized. Fitting several of these jobs into the individual nodes' schedules will necessarily leave gaps, since jobs have to wait until a sufficient number of nodes become available. The scheduler is "playing Tetris" with large, incompatible pieces, which can lead to under-utilization of the whole system. In contrast, cloud workloads typically consist of small tasks, each only using fractions of the available compute resources, but often in large quantities. Cloud computing offers flexible operating system and software environments which can be optimally leveraged in cases of embarrassingly parallel workloads that do not suffer from slow inter-node connections. A cloud environment offers more flexibility at the expense of the virtualization overhead and the loss of efficient, multi-node high-speed low-latency communication.

Large scale projects, e.g. the Large Hadron Collider from the particle physicist community, have a tendency to develop extremely large and complex software stacks. There exists a well-founded lay-assumption that a computational experiment that deals with objective values ought to be reproducible by other researchers. However this is often not the case as researchers are unaware of the details of their operating environment (operating system, application, and dependencies etc) [2]. One potential solution to the reproducibility issue is to be found in containerisation technology, which permits encapsulation of the entire operating environment. Unfortunately this approach has a couple costs; there is a modest performance penalty, and it requires that the container owner manages be the entire environment.

In addition to the more general problems of HPC and Cloud workloads, finding a truly optimal hybrid system should also consider resource allocation issues. An often encountered instance of this problem is when one parallel application on a HPC system requests 51% of the available resources while another application needs 50% of said resources. It is impossible to co-locate these workloads, which unfortunately leave a large portion of nodes idle if those are the only two tasks scheduled. Another resource allocation problem arises in the scheduling of jobs which different behaviours. In the case where two types of jobs (single-node, short term vs. multi-node, longer term) are submitted to a job scheduler with backfilling to maximise resource utilisation, the overall system can still experience sub-optimal utilization. The single node jobs, request fewer resources and will have priority over the multinode jobs, thereby reducing the cost-effectiveness of the expensive high-speed interconnect and resulting in idle cluster resources.

When looking for a hybrid system which offers flexibility and performance in the face of such problems, it is also worth noting that applications and their datasets usually have varied and characteristic computational workflows that are more or less appropriate for different computational architectures. It is certainly preferable from a user's perspective that a single system is capable of adapting to these diverse requirements, rather than requiring the user to migrate data between different systems depending on the task. A hybrid system must be found in order to address these sub-optimal resource utilization issues.

3 Hybrid Architectures

Both Melbourne and Freiburg Universities have been operating various compute clusters for more than two decades. While the variation in hardware architecture has decreased, new possibilities have arisen from X86 hardware virtualization becoming ubiquitous. It has allowed both sites to host multiple operating systems on a single server and to strictly separate their complete operating environments. Hardware and software stacks are decoupled. While widespread in computer center operation, virtualization is still underutilized in HPC environments. Projects like hybrid cluster approaches and the use of Virtualized Research Environments (VRE) drastically changes the landscape.

3.1 HPC with Compute Nodes as Cloud VMs

The University of Melbourne approach consists of a traditional HPC cluster with a high speed interconnect in one partition (or queue), and an alternative partition providing virtual machines (VM) managed through OpenStack as compute nodes. Using the Slurm Work-

load Manager multiple partitions are accessible; the bare metal partition in place for a traditional HPC architecture), while the National eResearch Collaboration Tools and Resources project (NeCTAR) research cloud provides generic virtual machines). Additionally there are private departmental partitions (the "water" and "ashley" partitions) , a specialist proteomics partition ("punim0095"), a general GPU partition, a large general-purpose graphics processing (GPGPU) partition for recipients of a specific Linkage, Infrastructure, Equipment, and Facilities (LIEF) grant from the Australian Research Council, and others. In total there are 24 partitions, including those used for debugging and testing purposes, and nodes can belong to multiple partitions simultaneously. Despite their heterogenous high level use cases, the general hardware specifications can be summarised as follows (Fig. 1):

- Physical partition: 12 core, Intel Xeon CPU E5-2643 v3, 3.40GHz, 256GB RAM.
- Cloud partition: 8 core, Intel Haswell, 2.3GHz, 64GB RAM.
- GPGPU partition: 24 core, Intel Xeon E5-2650 v4, 2.20GHz, 128GB RAM, 4 Tesla P100 cards.
- Ceph filesystem: 4.5PB for /home, /project, and /scratch storage.
- Network varies by partition; cloud partition on 10GbE with Cisco Nexus and physical partition with Mellanox ConnectX4 cards and SN2100 switch.

The VMs on the cloud partition use a common image just like the traditional HPC compute nodes, but with differing virtual hardware based on the results of job profiling and user requests. Each of these have an nodelist and are generated by VM images. Deployment of compute node according to partition is carried out with a simple script which invokes the OpenStack Nova service to deploy specific images. These can be deployed as either a static definition, or dynamically using Slurm's cloud-bursting capabilities. In addition the login and management nodes are also deployed as virtual machines.

Jobs are submitted to the Slurm workload manager, specifying which partition that they wish to operate on. The collection of virtual machines comes from the Melbourne share of the Australian-wide NeCTAR research cloud [3]. The VMs, with different virtual hardware, can be configured flexibly into different partitions in accordance with user needs. However, unlike a lot of VM deployments, overcommitment of resources is used. Early testing indicated that whilst there is good boundary separation through the virtualisation model, overcommit usage resulted in unexpected time mismatch errors on concurrent compute tasks. As a result the entire virtual machine architecture operates with a 1:1 ratio with physical machines. While this removes the normal advantages of overcommit for scheduling low utilization processes, it instead simplifies deployment and offers significant flexibility in extending or reducing the size of partition as required.

Of particular importance is assuring that the HPC "physical" partition has a high-speed interconnect. Mellanox 2100 switches with 16 x 100Gb ports with a mixture of 25/50/100Gb, maximum of 64 x 25Gb connections with RDMA over Ethernet and Cumulus Linux OS. An Message Passing Interface (MPI) "ping-pong" test was conducted between two compute nodes on separate infrastructure, with 40GbE RDMA over ethernet receiving better latency results than 56Gb Infiniband FDR14 on a comparable system [4].

3.2 HPC with Cloud VMs on Compute Nodes

The University of Freiburg runs an HPC cluster for Tier-3 users coming from different scientific communities and fields of interest. After a primary installation of 752 nodes in 2016 the cluster now commands more than 1000 compute nodes equipped with dual socket 10 core plus HT Intel Xeon E5-2650 v4, 2.20GHz CPUs and 128GB of memory each. The local operating system is remotely booted over the 1Gb Ethernet interface which also handles most of the user traffic. The cluster is equipped with 100Gb OmniPath offering 44 node islands which are aggregated by 4 100 Gb links. This high speed low latency network provides both MPI functionality for HPC jobs and access to the BeeGFS parallel filesystem.

Standard bare metal jobs are scheduled by Adaptive Moab. On top of the standard HPC jobs (bare metal jobs), it enables users to run virtual machines as standard compute jobs (VM jobs). In order to run VMs on a compute node, a virtualization hypervisor is installed on every compute node using the standard Linux Kernel-based Virtual Machine (KVM) hypervisor. This architecture enables users to run both bare metal and VM jobs on the same hardware, through the same resource manager, without partitioning the cluster into isolated parts. Users that require a special software environment and do not need direct access to hardware can use the VM jobs which provide Virtual Research Environments (VRE). A VRE in this instance is a container image with a complete software stack installed and configured by the user [5]. Thus, the file system of a virtual machine or VRE is a disk image presented as a single file. From the operator's perspective this image is a "black box" requiring no involvement, provisioning, updating, nor any other management effort. From the researcher's perspective the VRE is an individual virtual node whose operating system, applications and configurations as well as certain hardware-level parameters, e.g. CPU and RAM, can be configured fully autonomously by the researcher.

On typical HPC clusters the resource management is orchestrated through a scheduler. Since researchers using a hybrid compute resource are allowed both to submit jobs to the scheduler for bare metal computa-

tion and initiate VRE jobs, it is necessary that the scheduler is also responsible for the resources requested by virtual machines. If VMs are executed on the cluster without the knowledge of the scheduler, resources could get overbooked. A special workflow was developed to submit a request for a VM as a normal cluster job and let the scheduler handle the request for a new virtual machine. Thus, the hybrid HPC cluster concept becomes flexible and independent of the actual scheduler deployed.

The hybrid cluster setup is not without additional costs. For the orchestration of the virtual machines on the cluster the OpenStack framework is installed. If a compute job is designed to run in a virtual environment (VM) the Moab scheduler is configured to instantiate this VM through the OpenStack API. OpenStack is selected for its modular architecture to allow to choose the components needed for the cluster virtualization, such as network, image handling or a web interface and omit the rest. While introducing additional management and operational overheads, this allows building a virtualization environment for the cluster that meets the specific objectives. Four infrastructure nodes are dedicated to running the OpenStack services and preparing VM images. These nodes are not part of the compute resources available to the user.

3.3 Containerization

A further elaboration of the hybrid architecture is the use of containers in either aforementioned model. Whilst virtual machines simulate the hardware environment (even in a 1:1 environment), a container virtualizes at the operating system level. As containers can be imported with a exact level of consistency this is seen as solution to reproducibility issues in computation.

The most well-known application for containerisation is Docker. However Docker is not a good fit for the common tasks in research computing [6]. Docker is primarily designed for micro-services on an enterprise level, or during software development on local systems. Additionally, the Docker daemon runs as root, and the group with root privileges, which has security implications, e.g. when mounting file systems. These scaling and security issues leads HPC system administrators to be resistant to the installation of Docker. As an alternative, Singularity can be deployed without the issues of Docker, to obtain the benefits of containerisation [7]. Singularity is used at the University of Melbourne as it has good integration with the Slurm Workload Manager and MPI. Security issues are mitigated by ensuring that user privileges inside the system are the same as the privileges outside the system (i.e., as a user on an HPC) and that there are no root-owned daemons or root-level privileges for the group. Another container technology that also is considered appropriate for high performance compute environments is Shifter [8]. In either case, these containers can run on HPC nodes, whether they are bare-metal or virtualized, leading to the interesting situation of a virtualized operating environment (singularity) on a VM with its own operating environment, which is running on real hardware with an additional operating environment. The saying "that there is no cloud, there is just somebody else's computer" is doubly applicable for containers.

The following is a simple example to illustrate the use of the container within a Slurm job script that makes use of an existing container.

```
#!/bin/bash
#SBATCH --partition cloud
module load Singularity/2.4-GCC-6.2.0
singularity \
 exec vsoch-hello-world-master.simg \
 echo "Hello from inside my container!" \
 > output.txt
```

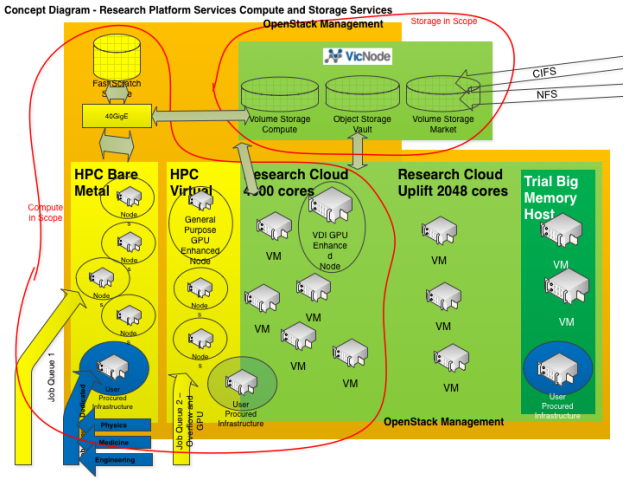


Figure 1: Concept diagram of the hybrid research platform services, combining bare metal and OpenStack resources at the University of Melbourne depicting the various partitions.

4 Workflow Design

Given the two competing primary compute resource variants, HPC scheduling and Cloud scheduling, it is necessary to orchestrate both systems through a mediator. In a hybrid approach one needs to define a primary scheduler which controls which jobs should run on which worker node and instruct the other scheduler to run within the boundaries of the other scheduler, especially respecting the scheduling decision of the other. The mediator could be a special service daemon which waits for requests from users and administrators and translates it to the OpenStack API and vice versa.

The University of Melbourne uses a traditional HPC workflow where job submission with Slurm Workload Manager occurs on different partitions according to whether they are based on physical or cloud architectures. At the University of Freiburg three workflows are present; job submission via Moab scheduler without running a resource manager client in the VM, job Submission via Moab scheduler with a resource manager client, The Terascale Open-source Resource and QUEue Manager (TORQUE) running in the VM, and job submission via OpenStack Dashboard/API.

4.1 Job Submission with Slurm Scheduler for different Partitions

A standard batch job submission workflow is a key part of the architecture. This follows the well-known path in HPC for job submission. As a unique feature, the placement of the virtual machine on the cluster nodes is scheduled by Moab and the job lifetime is coupled to the lifetime of the VM. This allows for a seamless integration with the jobs sent by other user groups and honors the fairshare policies of the cluster. A batch script makes resource requests on a particular partition. Based on the resource request, the information received from resource manager daemons, and the fairshare policies in place, the scheduler will allocate a time when the job will run which may change if other jobs finish early etc. At the appropriate time, the job will launch on one or more compute nodes, run, and write the relevant processed data and output to the directories as specified. The developed thin integration layer between OpenStack and Moab can be adapted to other batch servers and virtualization systems, making the concept also applicable for other cluster operators.

In the University of Melbourne model, the Slurm workload manager acts as the job scheduler and resource manager. Different partitions refer to the queues which a user may submit jobs and are varied by the physical or virtual hardware that have been provisioned (e.g., `#SBATCH --partition=cloud` or `#SBATCH --partition=physical`). For single node jobs, whether single or multi-core, a low speed (Ethernet) network and virtual machines are suitable, whereas for multinode jobs the physical partition with a high-speed interconnect is used. Thus, the physical architecture is optimised for the type of computational task required, increasing overall throughput and more efficiently allocating resource.

4.2 Job Submission via Moab Scheduler without running a Resource Manager Client in the VM

At the University of Freiburg, one use case is where users provide their VM images themselves. These VM images cannot be trusted and therefore they are not allowed to access cluster resources like parallel storage or user home directories. It is expected that the user is working with an external resource manager or is using the cloud-init procedure to start compute jobs within the VM. In this case the user's workflow is to submit a cluster job through the scheduler msub/qsub and, if resources are available, the job launches a virtual machine via the OpenStack API. When the VM boots data and compute instructions have to be injected by an external resource manager or through cloud-init into the virtual machine (possibly injecting the aforementioned cluster job script).

4.3 Job Submission via Moab Scheduler with a Resource Manager Client (TORQUE) running in the VM

A second use case at the University of Freiburg is when the user submits classic compute jobs to a different software environment on the cluster (Fig. 2, left branch). The software environment is represented by a VM in this case. This makes it necessary to install and run a TORQUE client in the virtual machine. In this use case the workflow begins with the user submitting a job through the scheduler (msub/qsub) with a specified image and job-script (e.g., msub -l image=¡id¿ job-script.sh). The job is then scheduled like any other bare metal job. If and when resources are available this job will trigger the start of a new virtual machine environment through the OpenStack API. When the virtual machine is booted the TORQUE client connects to the TORQUE server and receives the job which then is started within the VM. If this software environment is provided by an external source (user virtual machine) all necessary data has to be injected as well. This could be achieved for example with the stagein and stageout options from TORQUE.

4.4 Job Submission via OpenStack Dashboard/API

The third use case from the University of Freiburg is when the user submits compute jobs simply by creating a VM via the OpenStack web interface (Horizon) or OpenStack API (Fig. 2, right branch). These virtual machines then should be represented as a compute job in the Moab scheduler. The compute job script is injected via cloud-init into the virtual machine during boot and is executed in the virtual machine after the boot process is finished. In this use case the workflow is initiated by the user starting a virtual machine instance on the compute partition. OpenStack schedules this VM as any other bare metal job and, when and if resources are available, OpenStack will start up the virtual machine on the resources previously assigned by the Moab scheduler. When the virtual machine boots data and compute instructions have to be injected by an external resource manager or through cloud-init into the virtual machine.

4.5 Script and Data Injection

In the University of Melbourne case the injection of scripts and data to the virtual machine is not a significant issue, as the virtual machines are simply another node on the HPC system. The issues faced here are the same as other HPC environments. Primarily this means the need to firmly encourage users to manage their data so that it is physically close to the site of computation. Grace Hopper's famous reminder [9] to "mind your nanoseconds" is applicable even to many research scientists who ask whether remote data can have a mount-point on a HPC system in order to speed

up their computational time, and even across all compute nodes which are normally do not have access outside of the system network for security reasons. A more genuine issue is the access speed of shared mount points across a system, and education on the need for staging actions on local (e.g., `/var/local/tmp`) disk or faster shared storage (e.g., `/scratch`) before copying data to slower shared directories.

For Freiburg's architecture the situation is more complex. There are there possibilities to inject job scripts into a VM: cloud-init, TORQUE client and an external resource manager client.

- cloud-init: This is the easiest way to inject job scripts into a VM. A job script has to be specified when instantiating the VM. After booting the VM the script gets executed automatically.
- TORQUE Resource Manager Client: Since the Cluster uses the TORQUE resource manager new "virtual" clients can be added dynamically and so the job script will be injected directly through the resource manager to the client. The only challenge is to signal the TORQUE Client running the bare metal job not to execute the job script.
- External Resource Manager: If external scheduling and resource managers are implemented by a research group they can be used to inject job scripts into the VM. Once the virtual resource pops up after starting a VM on the bare metal cluster it registers at its external resource manager and then can be used as every other resource of this Resource Manager.

The nature of the VM image defines the difficulty of injecting data into the booted virtual environment. Usually virtual research environments are built by research groups containing the configuration and software which is necessary to solve the problems of that specific research groups. Since these VM images are externally built, the instantiated environments can't be trusted. In such environments mounting home and work directories is not allowed, so other methods for data injection and result retrieval from these environments must be available. There are two possibilities to do so:

- The job script can provide information on how to transfer data by copying or streaming it through the network
- The resource manager client features can be used to stage data into the VM and stage out the results. TORQUE supports staging data by specifying the options "stagein" and "stageout", the job output is copied to the remote site automatically if not specified otherwise in the TORQUE client config ($usecp).

For trusted virtual environments provided by cluster administrators, these issues are not present. For example replicating a modules software environment for experimental particle physicists. These virtual environments are fully controlled by the same people running the bare metal environment and so can be trusted. In these environments, home and work directories can be mounted, and users can be identified and authorized through the same mechanisms as in the underlying bare metal environment. They can be simply leveraged by users as special software environments with no difference to the bare metal system (ignoring the fact, that it wouldn't support multi-node usage).

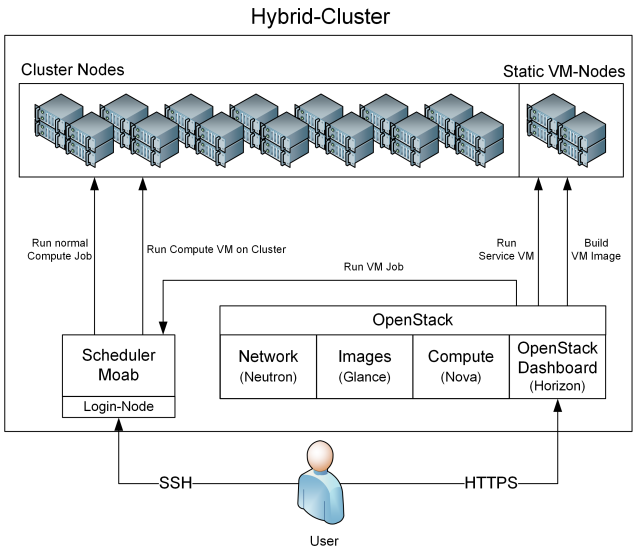


Figure 2: The Freiburg HPC cluster user has multiple options to submit compute jobs, either via the traditional way submitting it to the scheduler (left branch) or interactively by starting an appropriate virtual machine containing the relevant workflow (right branch).

5 Conclusions

Hybrid clusters make research groups more independent from the base software environment defined by the cluster operators. Operating VREs brings additional advantages like scientific reproducibility but may introduce caveats like lost cycles or a layered job scheduling. However, the advantages of this approach makes HPC resources more easily available for broader scientific communities. Both models are meant to be easily extensible by additional resources brought in by further research groups.

The two models - HPC with Cloud VMs on Compute Nodes, and HPC with Compute Nodes as Cloud VMs - represent different hybrid systems to solve different problems. In effect, the University of Freiburg model provides a "cyborg", where the HPC compute nodes are replaced with cloud virtual machines, whereas the University of Melbourne model provides a "chimera", a multi-headed beast where the virtual machines have become new cloud nodes. In the former case there was a desire to make existing compute nodes available to researchers for their particular configurations. In the latter case there was a desire to make virtual machines accessible to an HPC system to increase cost efficiency and improve throughput.

The two approaches illustrate the importance of HPC-Cloud hybrids in general purpose research computing.

Migrating complex and diverse software stacks to new HPC systems every few years constitutes a major effort in terms of human resources. By virtualization of these complex software environments, one sacrifices a fraction of performance, but one gains the possibility to run these complex software environments in cloud systems and thus literally anytime, anywhere and on any scale. This flexibility, in many cases, outweighs the loss in performance [10, 11, 12].

For future developments, the University of Melbourne plans to extend their model to provide the ability to include cloud bursting to external providers (e.g., Amazon, Azure), and hosting highly varied X86 configurations on the same system, although this requires equivalent real or virtual hardware, partial replication of the operating environment, and accounting for latency between the local and remote sites. A further option of HPC cloud-bursting that will be subject to further investigation is Moab/NODUS cloud-bursting by Adaptive Computing. The principles used by NODUS are similar to the exploration here; a workload queue has an elastic trigger that uses an API key to an template image which then invokes nodes, deploys upon them and potentially cluster nodes simultaneously, completes the job, and transfers the data as necessary. The ability for fine-grained task parallel workloads across local HPC and remote cloud providers suggests an obvious difficulty.

The University of Freiburg model is rather complex and took a while to mature. It is in production since the official start of the cluster in mid 2016. The HPC/cloud group hopes to improve their model, mapping Moab commands to OpenStack commands allowing to pause/hibernate and resume the virtual machine for preemption or maintenance instead of killing a job. In addition the possibility of mapping the live migration of virtual machine to Moab during runtime will give the opportunity to migrate compute jobs during runtime to optimize the overall cluster utilization. For the next generation HPC cluster we will reconsider the options to reduce the complexity of the VRE scheduling. Following the Melbourne model a distinct cloud partition for the future system is definitely an option.

Acknowledgment Lev Lafayette would like to thank Bernard Meade, Linh Vu, Greg Sauter, and David Perry from the University of Melbourne for their contributions to this document. The authors from Freiburg University would like to thank the Ministry of Science, Research and the Arts of Baden-Württemberg (MWK) and the German Research Foundation (DFG) which funded both the research infrastructure NEMO and the ViCE eScience project on Virtual Research Environments.

References

[1] L. Lafayette and B. Wiebelt, "Spartan and NEMO: Two HPC-cloud hybrid implementations," in *13th International Conference on e-Science*, IEEE, oct 2017.

[2] V. V. Sochat, C. J. Prybol, and G. M. Kurtzer, "Enhancing reproducibility in scientific computing: Metrics and registry for singularity containers," *PLOS ONE*, vol. 12, p. e0188511, nov 2017.

[3] Z. Li, R. Ranjan, L. O'Brien, H. Zhang, M. A. Babar, A. Y. Zomaya, and L. Wang, "On the communication variability analysis of the NeCTAR research cloud system," *IEEE Systems Journal*, vol. 12, pp. 1506–1517, jun 2018.

[4] B. Meade, L. Lafayette, G. Sauter, and D. Tosello, "Spartan HPC-Cloud Hybrid: Delivering performance and flexibility," 2017.

[5] M. Janczyk, B. Wiebelt, and D. von Suchodoletz, "Virtualized research environments on the bwForCluster NEMO," in *Proceedings of the 4th bwHPC Symposium*, 2017.

[6] G. Kurtzer, "Singularity: Containers for science, reproducibility, and high performance computing." Keynote at Stanford Conference, 2017.

[7] G. M. Kurtzer, V. Sochat, and M. W. Bauer, "Singularity: Scientific containers for mobility of compute," *PLOS ONE*, vol. 12, p. e0177459, may 2017.

[8] D. M. Jacobsen and R. S. Canon, "Contain this, unleashing docker for HPC," in *Proceedings of the Cray User Group*, 2015.

[9] L. Gilbert, *Particular Passions: Grace Murray Hopper (Women of Wisdom)*. Lynn Gilbert Inc., 2012.

[10] A. Gupta and D. Milojicic, "Evaluation of HPC applications on cloud," in *Sixth Open Cirrus Summit*, IEEE, oct 2011.

[11] P. Church and A. Goscinski, "Iaas clouds vs. clusters for hpc: a performance study," in *Cloud Computing 2011 : The 2nd International Conference on Cloud Computing, GRIDS, and Virtualization*, pp. 39–45, IARIA, 2011.

[12] A. Marathe, R. Harris, D. K. Lowenthal, B. R. de Supinski, B. Rountree, M. Schulz, and X. Yuan, "A comparative study of high-performance computing on the cloud," in *Proceedings of the 22nd international symposium on High-performance parallel and distributed computing - HPDC '13*, ACM Press, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 06-15 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Harmonic Sensitivity Analysis of Power System with Wind Power

Jignesh Patel*, Satish Joshi

Department of Electrical Engineering, The Maharaja Sayajirao University of Baroda, Vadodara - 390002, India

ARTICLE INFO

Article history:
Received: 29 June, 2018
Accepted: 25 August, 2018
Online: 08 September, 2018

Keywords:
Harmonic Resonance
Modal Analysis
Eigenvalue
Modal Sensitivity
Eigenvector

ABSTRACT

As the contribution from renewable energy sources increase, the power quality degradation has become a challenge for the power system operator. There are major two factors causing power quality deterioration. First is the current harmonic generation by the power electronic converters used in renewable energy sources, and second is the magnification of harmonic voltage due to harmonic resonance condition educed by power factor correction capacitors. The motivation behind this work is power quality related problems, observed recently in the field. To analyze the problem a comprehensive approach is adopted. A practical wind farm configuration is considered for study. Using an eigenvalue analysis method, critical resonance modes are determined for different value of capacitor connected to different buses. To check the effect of change in network parameter on eigenvalue, the sensitivity analysis is carried out and how resonant point is varied with the capacitor value and connection point is discussed by simulating different cases. The results are useful in determining the filter value and location for maximizing its effectiveness in curbing the harmonic problems.

1. Introduction

In the recent years, several problems of Wind Generator tripping due to over-voltage is observed. When the problems were investigated, it is found that the harmonic current, generated by the wind generator converter, produces high voltage due to harmonic resonance. Also, several incidences have been reported of converter tripping due to over-voltage caused by the harmonics. This is the prime reason of carrying out this work to know the exact problem and finally find the applicable solution. In the event of harmonic resonance, the harmonic impedance becomes very high four harmonics of frequencies near to resonance point and theoretically become infinite at the resonance frequency. Under the resonance condition, even a small amount of harmonic current is enough to produce high harmonic voltage drop, which is ultimately resulting in overall voltage distortion.

There are various methods to analyze the harmonic problem. Generally, the frequency scan method used to locate the harmonic resonance point. Though, it is simple and easy to use, but it does not give information about the participation of different network components and also how the resonant point is affected by changes in the value of the components.

*Corresponding Author: Jignesh Patel, Email: jignesh.pravinbhai@gmail.com

To analyze the problem, several techniques were investigated. Finally, based on the ease of use and potential of in-depth analysis, the modal analysis is adopted for analyzing the problem in this work.

In conventional power generation, harmonics is virtually not generated by the generators. Also, the resonance condition does not prevail on the generator bus because of absence of shunt capacitors. But, both of these are present in the Wind Farm. It may impact significantly, if the resonance point coincides with the one of the frequencies of generating harmonic currents.

Harmonic Resonance can become complicated, because of complex network configuration and presence of multiple harmonic sources in the network. A power system network contains many inductive, capacitive and resistive elements. An interaction among inductive and capacitive elements leads to oscillations. The period of oscillation depends on the resistive component. Phenomena. Though, the resistive elements cause the loss of power, but they also act as a damping element, which reduces the harmonic magnitude exponentially and supports the harmonic stability. The exchange of energy between capacitive and inductive elements appears as an oscillations, the frequency of oscillation depends on the value of participating capacitor and inductor value. Such The

The network is characterized by the critical frequency known as resonance frequency. If such a network is excited by the source having resonance frequency, will lead to a sustained oscillation.

Traditionally, the frequency scan analysis is a primary tool to find out the resonance in the system, but this technique lacks of follow information.

1. Which combination of L & C are involved in the resonance?
2. How many modes are participating in a given resonance?
3. What is the sensitivity of a resonance w.r.t different network element/component?
4. The effect of resonance on different buses of the network?
5. Which is the most effective way to avoid the harmonic resonance?

To answer these questions, a survey has been made and suitable analysis is selected, which is presented in this work.

All wind parks are unique in characteristics. They all have resonance frequencies that are dependent on various factors like grid topology, connected generators, and reactive power equipment used. Furthermore, the impedance and the resonance points of Wind Park changes with the change in number of turbine and capacitor or filter banks in operation.

The connecting cable has also a predominant effect on the resonance frequency. The internal cable connections in a Wind Park connect the turbine of the wind park, whereas the transmission cable connects the wind park to the nearby substation [1]. The total length of the cable varies from a few kilometers to few tens of kilometers. The length of cable, and hence the cable capacitance plays an important role in harmonic resonance.

Capacitor banks and reactors presence also increases the possibility of resonance in the wind power plant. Several capacitors are connected in the form of bank to improve the power factor and reduce the reactive power absorption from grid [2]. The capacitor bank at the PCC level is also used to support the voltage in abnormal conditions. There may be shunted reactors connected to the transmission cable terminations to compensate the high capacitance of a cable. All these components create conducive conditions for harmonic resonance.

In this work, a systematic approach is presented. Starting with the admittance matrix formation, followed by the finding of eigenvalues of discrete frequency points. For resonance or critical frequency, eigenvalue becomes very low or zero. These eigenvalues are then analyzed further to find out the effect of change in a network element on this eigenvalue. Because, the network configuration keeps on changing, depends on the operational requirement and so is the eigenvalue and critical frequency. So, it becomes imperative to find out, how firm or steady is the any resonance frequency.

The structure of this work is explained here. First, the contemporary literature is surveyed in section - 2. The modelling of component is given in section - 3, succeeded by the section - 4, explains the stepwise method for determining resonance frequency. The sensitivity analysis is briefly presented in section - 5. The problem is defined in section - 6 and different cases are elaborated from section - 7 to 10. The filter design to reduce the effect of harmonic resonance is depicted in Section -11. Finally, the work is concluded in section - 12.

2. Review of Related Literature

Various literatures are investigated to find out the work done in the field wind power effect on power quality. There are quite a few numbers of literatures are available, which has focused primarily on the power quality aspect of the wind power. In reference [3], harmonic emission from four different wind park is measured and compared to the harmonic spectrum. Further, the harmonic emission from the wind power converter is small as compared to any harmonic emission load. And mostly inter-harmonics are generated by wind converters. The most important thing found in this work is the frequency - duration plot, depicts how the spectrum is over a period of time. The significant change in the spectrum is observed during the measurement period, not only in magnitude, but also the ratio between different components. It is interesting to know from this work that, some components of harmonics are co-related with the active power variation, but others are not. This is very confusing. Finally, the work is concluded with the comment that, a measurement of one location during a short period of time cannot be represented as the harmonic emission due to wind power installation.
In reference [4], measurement of voltage and current distortion, at two locations inside the wind farm, is compared. The variation of the spectrum with respect to time is presented to show the emission in a Wind Park.

Harmonic variation with the wind farm operating point and the random characteristics of their magnitude and phase angle is measured extensively and the deterministic and stochastic characterization is analyzed in reference [5]. Similar work is found in reference [6]. In this work, the author has done extensive field measurements on commercial available wind turbines and revealed several important observations. In the opinion of this work, the spectrum of harmonic from a wind turbine is extended to few kHz, but with the increase in size of WT, the spectrum limited to low frequency, around less than 1 kHz. Among the several other observations, the important observation is about the presence of lower frequency harmonics (5th, 7th etc…). The lower order harmonics are generally not expected from PWM converter. The lower order harmonics follows the Normal distribution, whereas Rayleigh distribution is found in higher order harmonics. Further, the low order harmonics synchronized with fundamental frequency, but high order harmonics varies randomly between 0 to 2π.

A structured framework is presented in reference [7] to analyse the harmonics emission and the resonance condition in the wind power interconnection with the grid. According to this work, the wideband harmonics from wind turbines are stochastic in nature and they are associated with the active power production. They may adversely interact with the grid impedance and cause unexpected harmonic resonance. This issue needs to be addressed

comprehensively at the planning stage, should become more critical as wind power contribution increases in the grid. Unlike conventional power generation, wind power is distributed in nature. So, it is very complex when it comes to analysing the harmonic generation from wind converters. To address this issue, an approach of aggregation of harmonics from individual turbine using a method, similar methodology of IEC 61400 - 21, is presented in reference [8]. In this work, the wind farm with different layout is simulated. Then this result is compared with the aggregated harmonics calculated using proposed methodology. The result presented in this work suggests the use of the proposed methodology for summation of harmonics from the wind farm. According to this work, for better result, both, the accurate modelling of wind farm topology and supply system is very much essential. Similar to this work is found in reference [9], where power quality assessment is done as per IEC 61400-21 procedure.

Another approach of aggregation of wind farm for harmonic propagation is reported in [10]. In this work, detailed model of DFIG is used and using this an aggregated DFIG model is developed to study the harmonic propagation. Frequency domain approach is taken to study the problem. The author emphasizes that all the elements of a system are important, because they are frequency dependent and the whole system can affect the resonance frequencies. The analysis part of this work showed that increase in number of aggregated DFIG cause a shift in the harmonic resonance order.

Reference [11] presented the network reduction method for reducing detailed low - voltage wind farm network to assess the effects of its connection onto a main interconnected transmission system. Application of this method reduces the number of network to be modelled and number of injection nodes which reduces the study time. To demonstrate and prove its validity and effectiveness, the results are compared with the IEC current aggregation method.

Harmonic problems become a concern with the growing number of Wind Power Plant [12], because along with a non-linear load, generation side also contributes to the harmonic emission. In Wind Farm, several conditions may give rise to the resonance phenomenon that will amplify the effect of the harmonic frequency. Add on to this, overview of international standards, grid codes, and basic modelling of Wind Power Plant is also given in this literature.

Reference [13] discusses how a wind power installation can affect the harmonic level in a number of ways. According to the author, Wind Turbines are an additional source of harmonics, particularly non-characteristics harmonics. The resonance point is also depends on strength of the grid. Stronger grid will have a higher order of harmonics, whereas weaker grid gives rise to a lower order of harmonics. Further, the author has extensively explained the sources of primary distortion and secondary distortion. Also, how increase in no of wind power installation weakens the grid is elaborated.

3. Modelling of Components

Electrical components are represented by mathematical equations for analysis purpose. The validity of mathematical equation depends on many factors and it is applicable for certain operating range only. Also, the method of modelling affects the outcome of the analysis. Here, by taking the case of cable modelling, an effort is made to explain the associated difficulties.

HV Cables and Transmission lines are represented either by equivalent PI or T circuit. How the resonant frequency of two different models of the same component is differing, is explained here by example.

A cable with following parameters is modelled.

R_{SYS}=0.04 Ohm
R_{CABLE} = 0.835 Ohm (0.13 Ohm/km)
L_{CABLE} =12.73 mH (0.112 Ohm/km)
C_{CABLE}=4.2 uF (12.73 Ohm/km)

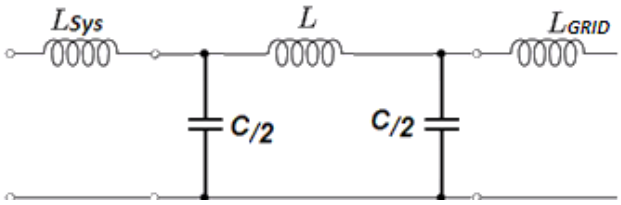


Figure 1: 'PI' model of Cable

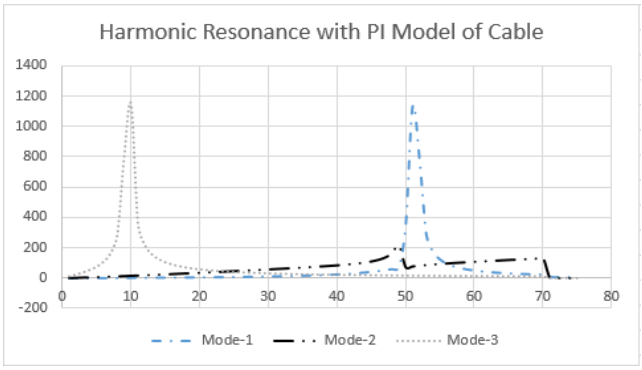


Figure 2: Harmonic Resonance with 'PI' model of Cable

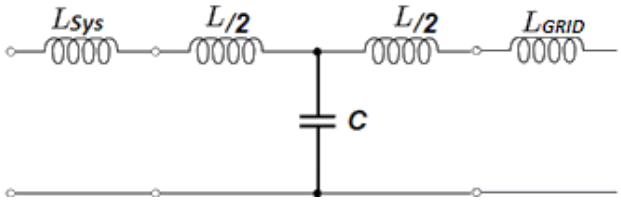


Figure 3: 'T' model of Cable

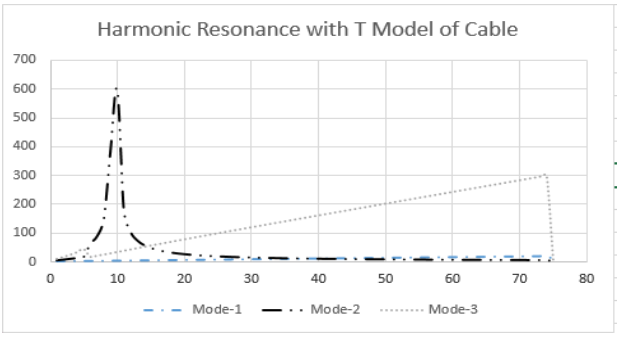


Figure 4: Harmonic Resonance with 'T' model of Cable

Comparison of figure 2 and 4 reveals the effect of modelling on the resonance mode. Mode 1 and Mode 3 are predominant with 'PI' model, whereas Mode 2 is significant and visible in 'T' model of cable. It is important to note that the Mode - 2 with PI model having resonance frequency near to the Mode - 1 of 'T' network. Thus, it is inferred from the result that, the mode number depends on the number of capacitive elements in the system.

In 'PI' model, the cable capacitance is split into two halves and represented in the form π configuration, so there are two modes one at around 10th order and second is around 50th order. Whereas, in the case of 'T' model, only one mode around 10th order is significant, because of lumping of cable capacitance and represented as one element in the form of 'T'.

The nodal impedance or self-impedance of a network is given by

$$Z_{11} = \lambda_1^{-1} PF_{11} + \lambda_2^{-1} PF_{12} + \lambda_3^{-1} PF_{13}$$

$$Z_{22} = \lambda_1^{-1} PF_{21} + \lambda_2^{-1} PF_{22} + \lambda_3^{-1} PF_{23}$$

$$Z_{33} = \lambda_1^{-1} PF_{31} + \lambda_2^{-1} PF_{32} + \lambda_3^{-1} PF_{33}$$

The effective impedance at any bus is depends on participation factor and modal impedance. Mode with smaller values or poor participation factor, may be omitted for calculation of bus impedance. So, when the impedance at any bus is calculated using any method, the result will not differ much. This is because of change of bus participation factor for a given mode. This can be shown here by calculation, but it is avoided to maintain the privity.

4. Determination of Resonant Frequency Based on Critical Mode

Generally, frequency scan is used for finding harmonic resonance. Though, it is an effective method, but lacking in some aspects. For e.g. frequency scan doesn't find the involvement of passive elements for a particular resonance frequency. To overcome, this difficulties, modal analysis is adopted in this work. The modal analysis overcomes the points, in which, frequency scan lacks. Here the modal analysis is explained in detailed.

The first step of analysis is the formation of 'Y' matrix.

$$[Y_h] = \begin{bmatrix} Y_{11}^h & Y_{12}^h & Y_{13}^h \\ Y_{21}^h & Y_{22}^h & Y_{23}^h \\ Y_{31}^h & Y_{32}^h & Y_{33}^h \end{bmatrix} \quad (1)$$

The voltage at different buses is given by

$$[V_h] = [Y_h]^{-1}[I_h] \quad (2)$$

$[Y_h]$ is the network admittance matrix at harmonic order h. $[V_h]$ is the voltage at different nodes, and $[I_h]$ is the harmonic current injected at a different nodes. The admittance matrix is a function of network configuration and the frequency. Harmonic sources having frequencies near to the resonance frequency, the admittance matrix offers very high impedance and at resonance frequency the admittance matrix becomes singular. Some nodal voltages rises sharply, when $[Y_h]$ matric approaches singularity. The transition of admittance matrix to singularity offers an effective way to analyze the harmonic problem. The eigenvalue analysis has been used for more than a decade to analyze such problems. The eigenvalue analysis technique is applied to system of linear differential equations. In fact they are applicable to any system involving matrices. In power system, it has been widely used to analyze the dynamic stability.

According to [3], the matrices can be decomposed in to following form

$$[A] = [L][\lambda][T] \quad (3)$$

Where, $[\lambda]$ is the diagonal matrix contains eigenvalues, and $[L]$ and $[T]$ are the left and right eigenvectors respectively. The product of left and right eigenvector is identity matrix, applying this to admittance matrix,
So,

$$[L][T] = [I_n]$$

$$[V] = [L][\lambda]^{-1}[T][I] \quad (4)$$

Multiplying both side by$[T]$,

$$[T][V] = [T][L][\lambda]^{-1}[T][I]$$

$$[T][V] = [\lambda]^{-1}[T][I] \quad (5)$$

Defining, $[U_m] = [T][V]$ as the modal voltage vector and $[J_m] = [T][I]$ as the modal current vector respectively, the above equation can be written as

$$[U_m] = [\lambda]^{-1}[J_m] \quad (6)$$

The inverse of the eigenvalue, $[\lambda]^{-1}$has the unit of impedance and is known as the modal impedance$[Z_m]$. From (4), it can be seen that, if λ_i=0 or is very small, even a small injection of modal current J_i will lead to a large modal voltage V_i.

$$[T][I] = [J_m] \quad (7)$$

Nodal voltage,

$$[T][V] = [U_m] \quad (8)$$

And nodal current,

$$[T][I] = [J_m] \quad (9)$$

Determination of system resonant frequencies using finding of critical mode is possible by finding out the eigenvalue of 'Y'

matrix at various discrete frequency. System resonant frequencies can be located from Admittance-frequency curves of the modal Admittance, where admittance approaches vary low value, near to zero. The procedure of resonance determination by using modal analysis is illustrated in flow chart given in figure 1. The maximum magnitude of modal impedances at a resonance can be interpreted as the degree of resonance severity.

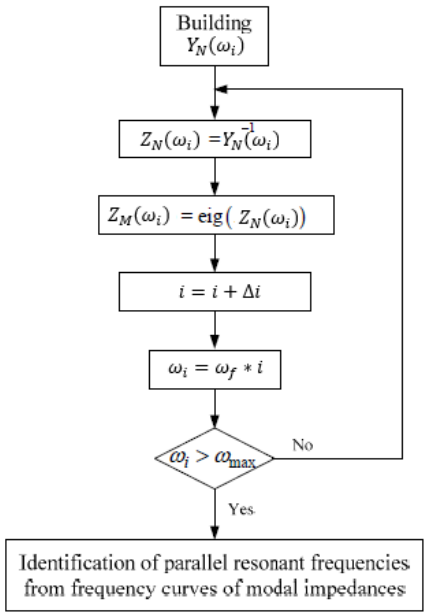


Figure 5: Flow Chart for Critical Eigenvalue Finding

Plot impedance-frequency curve and identify the peaks from curve. Now reduce theΔi, and run the scan again between two frequency value, where the peaks are observed. Stop the scan when$i < i_{min}$.

5. Modal Sensitivity Analysis

The modes or eigenvalues varies with the passive elements of the circuit. So, as the element value changes, mode value also change. The sensitivity of any mode with respect to the any element of matrix is a function of left and right eigenvector [3-5].

According the reference [6], the modal sensitivity is given by equation 10. First order eigenvalue sensitivity of λ_k w.r.t parameter a_{ij} is given by

$$\frac{\partial \lambda_k}{\partial a_{ij}} = l_i^k r_j^k \ (i, j, k = 1,2,3 \ldots . n) \tag{10}$$

Where,
l_i^k = Left eigenvector
r_j^k= right eigenvector

Integration of Wind Farms to the power system has some challenges, one of them is the harmonic resonance, which result in to magnification of harmonics. Several tripping has been observed in the recent days in Wind Farms. The problem analysis pointed out the harmonic resonance as a root cause. Before inverters are connected to the grid, a detailed assessment is required to find out network impedance and also the behavior of network impedance with the change in frequency and also with the change in passive elements. Direct connection of inverters to the grid may give rise to power quality problems. One such problem is analyzed here.

The harmonic emission of WT converters are normally low, and hence voltage distortion usually remain below standard limits. However, the presence of parallel resonance in the Wind Park grid may increase the voltage distortion above the limits and also affect WP harmonic emission to the main grid. Several work analyze the resonance problem at WT terminal by frequency scan method. But as mentioned in section – 4, it is lacking in some crucial aspects and does not provide information related to origin of harmonic magnification. It is used widely because it requires less computation efforts and give quick overview of harmonic resonance. So, the frequency scan method can be used as a compliment of modal analysis.

The wind farm under study consists of 60 WT of 2 MW each. The aggregate capacity of the wind farm is 120 MW. Each WT is connected to the 0.440 /1.1 KV Transformer. The collector level, which is known as pooling substation, is at 33 KV. At the Pooling substation, the voltage is raised through 1.1/33 KV, 12 MVA transformers. The polling station is connected to the 132 KV transmission substation through 36 KV, 485 Sq.mm, and Armoured XLPE cable of 10 km in length. The power is then evacuated from 132 KV substation through transmission line with Moose Conductor.

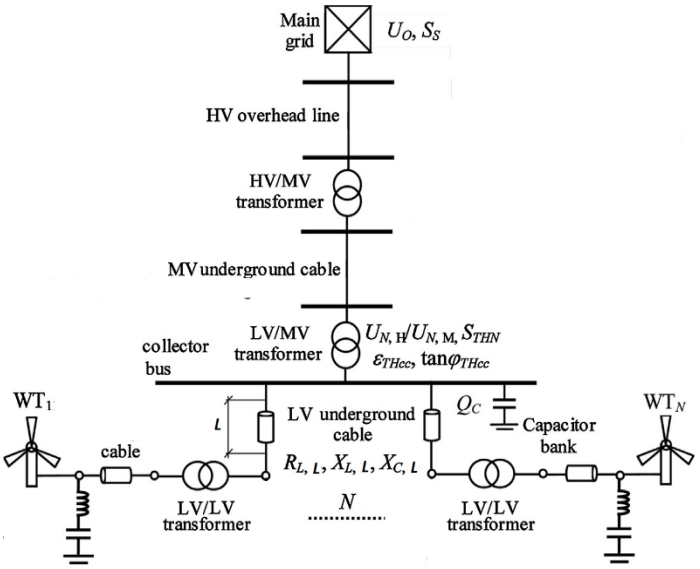


Figure 6: Network Diagram

While running in parallel to the grid, each inverter is able to produce up to 1.95 MW of power. During moderate wind speed condition, each WT generates around 1.2 – 1.3 MW of Power. There was intermittent inverter tripping observed from the site. The inverter tripping history recorded the "Line Overvoltage Fast". This implies that the line voltage has exceeded 120% for more than

8 cycles (i.e. 120 ms). Also, there was an unusual hissing sound before the tripping.

7. Case – I: Single Wind Turbine with 200 kVAr Capacitor at the Terminal of Wind Turbine

In this case, only one turbine is considered in the Wind Park and 200 kVAr capacitor is connected to its terminal for PF Correction.

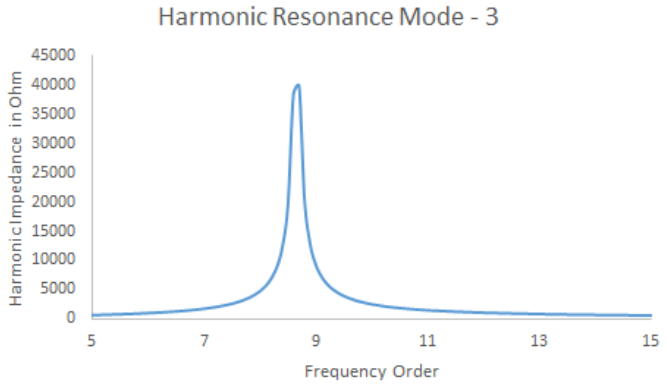


Figure 7: Harmonic Mode – 3, 9th Order

The network is further analyzed to find out the effect on the concerned bus. Using analytical method, the resonant frequency is found out. The first dominant frequency found is 435 Hz, which is near to 9th order. The PF correction capacitor along with combined reactance of LV transformer (0.440 / 1.1 kV), reactance of LV cable and reactance of LV/MV (1.1/33 kV) transformer, forming a parallel resonance at 11th order. PF Capacitor value is 0.595 uF and transformer reactance is 43.56 Ohm.

$$h_{res} = \frac{1}{2\pi f}\sqrt[2]{\frac{1}{(0.595e^{-6})(0.2236)}} = 8.72 \tag{12}$$

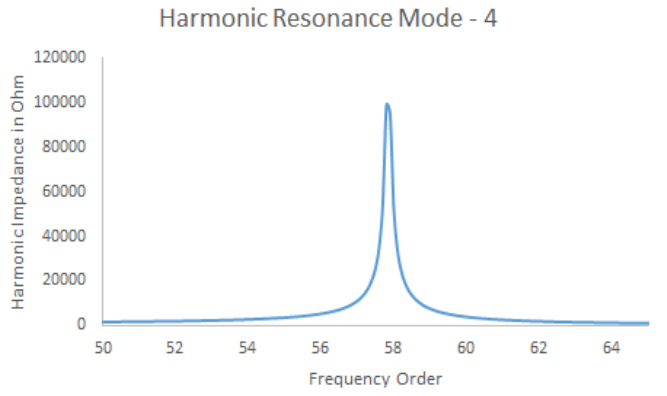
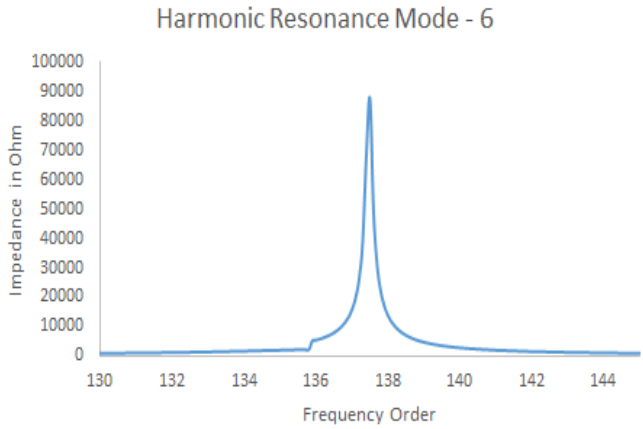


Figure 8: Harmonic Mode – 4, 59th Order

Harmonic Resonance Mode – 4 is contributed by 33 kV Cable capacitance of 0.36 uF. Cable capacitance is acting along with 33 kV Cable reactance, 132 kV transformer reactance, 132 kV line reactance and Grid reactance, having combined value of 2.40 Ohm. According to analytical expression the harmonic frequency order is 60.7th order as given in equation 13. This is very near to the mode-4 as shown in figure -7.

$$h_{res} = \frac{1}{2\pi f}\sqrt[2]{\frac{1}{(0.36e^{-6})(7.63e^{-3})}} = 60.7 \tag{13}$$



The Mode – 6 is also contributed by 33 kV Cable capacitance. As, the cable is represented by PI equivalent circuit, the total capacitance is divided in to two halves and each capacitor is lumped on two sides of the cable impedance. So, according to analytical expression, the cable capacitance of 0.36 uF is reacting with effective reactance of 33kV Cable, 132 kV transformer, 132 kV line and Grid, having aggregate value of 0.533 Ohm (1.6 || 0.8). Resonance frequency as per the analytical equation (14) is 128.8th order, which is near to the Mode – 6.

$$h_{res} = \frac{1}{2\pi f}\sqrt[2]{\frac{1}{(0.36e^{-6})(1.7e^{-3})}} = 128.8 \tag{14}$$

Table 2: Eigenvalue Sensitivity for Case-I

Modal Sensitivity	Bus-1	Bus-2	Bus-3	Bus-4	Bus-5	Bus-6	Bus-7
Mode-3	0	0	0.0197	0.2411	0.6227	0.1128	0.0039
Mode-4	0	0	0.0113	0.4139	0.1736	0.1556	0.2456
Mode-6	0.5	0	0.0103	0.0045	0.0124	0.1503	0.3224

In this case, number of wind turbine is increased to two. It can be seen from result plot in figure 8. Two resonance points is seen, one is at 8.1th and second is at 9.5th order. The reason for two resonance peak is multiple capacitors, one is due to the PF capacitor and second is due to the LV cable capacitance.

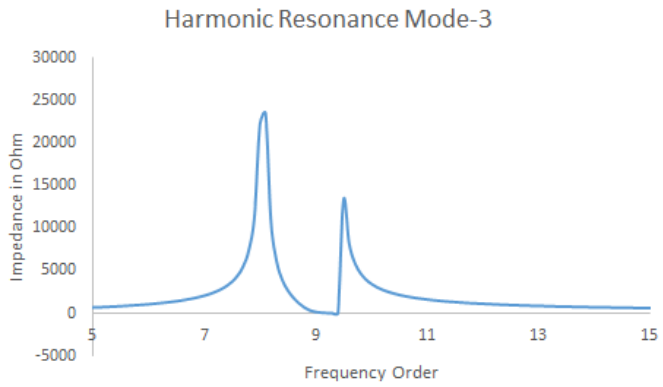


Figure 10: Harmonic Mode – 3, 8.1th and 9.5th Order

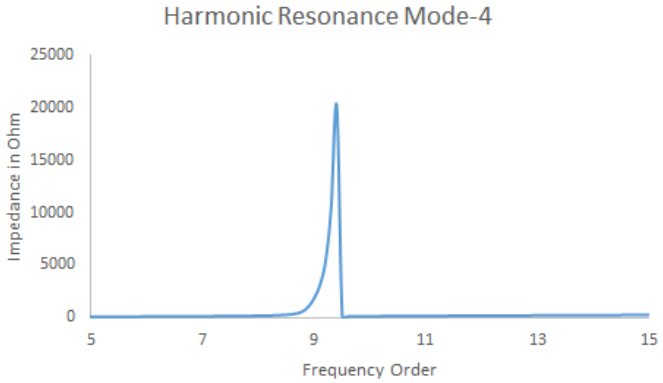


Figure 11: Harmonic Mode – 4, 9.4th Order

Mode – 4 is near to the second peak of Mode – 3. The frequency span between two resonances depends on the impedance between two capacitors. Here, the 1.1kV cable impedance parting cable in to two halves, in PI form. Due to this, two peaks are observed. As the cable impedance is very low, two peaks are observed nearby.

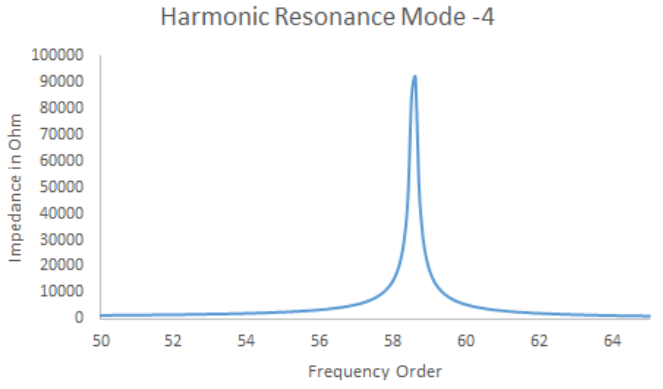


Figure 12: Harmonic Mode – 4, 58.6th Order

Mode – 4 has second peak at 58.6th order, which is the same as Mode – 4 observed in Case – I. This remains unaffected, as the addition of second wind turbine is not affecting the network above 1.1 kV collector bus to Grid. The same is observed in Mode – 6, as shown in figure – 11. This is also not affected by addition of additional wind turbine.

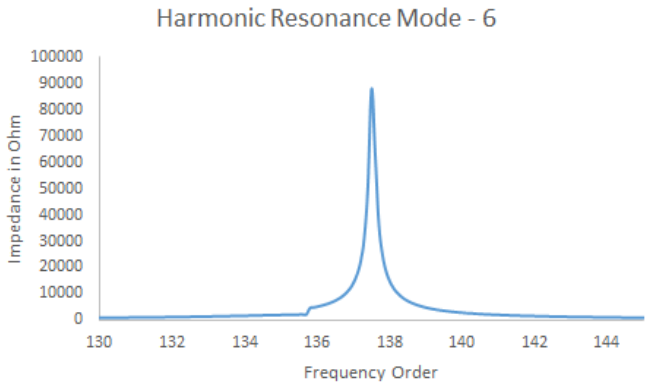
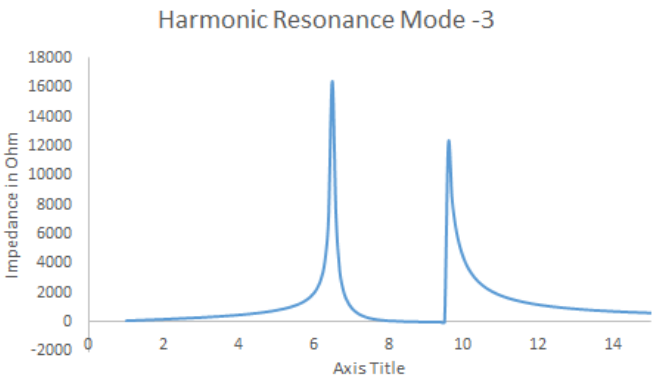


Figure 13: Harmonic Mode – 6, 137.5th Order

Table 3: Eigenvalue Sensitivity for Case-II

Modal Sensitivity	Bus-1	Bus-2	Bus-3	Bus-4	Bus-5	Bus-6	Bus-7
Mode-3 (8.1)	0	0	0.0563	0.3751	0.4794	0.0859	0.0036
Mode-3 (9.5)	0	0	0	0.0059	1	0.0382	0.01
Mode-4	0	0.0131	0	0	0.0284	0.6769	0.333
Mode-6	0.5	0	0.0103	0.0042	0.0105	0.1526	0.3224

9. Case – III: Six Wind Turbine with 200 kVAr Capacitor at the Terminal of Wind Turbine



In this case, Mode – 4 and Mode – 6 remains unchanged, as this are network dependent modes and as the network configuration is not changed, the modes is also remaining unaffected. The resonance sensitivity analysis is given in Table – 4.

Table 4: Eigenvalue Sensitivity for Case-III

Modal Sensitivity	Bus-1	Bus-2	Bus-3	Bus-4	Bus-5	Bus-6	Bus-7
Mode-3 (6.5)	0	0	0.1548	0.3946	0.3747	0.0731	0.0034
Mode-3 (9.6)	0	0	0	0	0.9966	0.0031	0.0046

10. Case – IV: Single Wind Turbine with 200 kVAr Capacitor at the 1.1 kV Collector Bus (Bus – 3)

This case is same as Case – I, except one change i.e. PF Capacitor is connected at Bus – 3 instead of Bus – 1. This case is analyzed to study the effect of location of PF Capacitor. Figure – 13. It is observed that, the resonance frequency is relocated to 21.7th order.

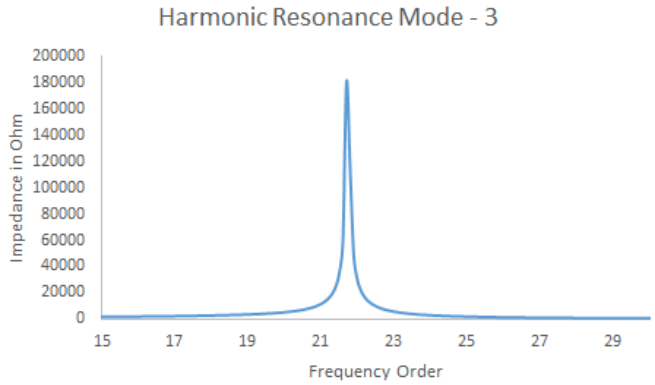


Figure 15: Harmonic Mode – 3, 21.7th Order

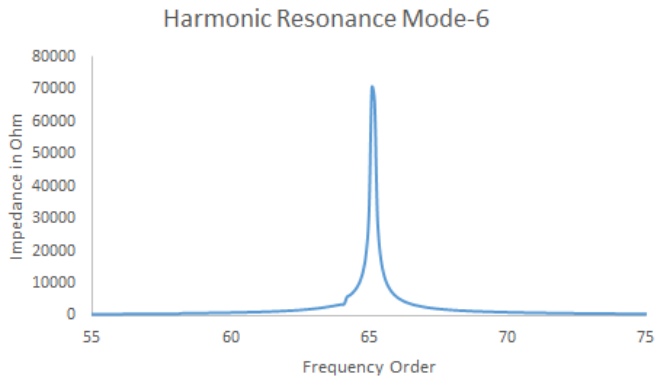


Figure 16: Harmonic Mode – 6, 65.2nd Order

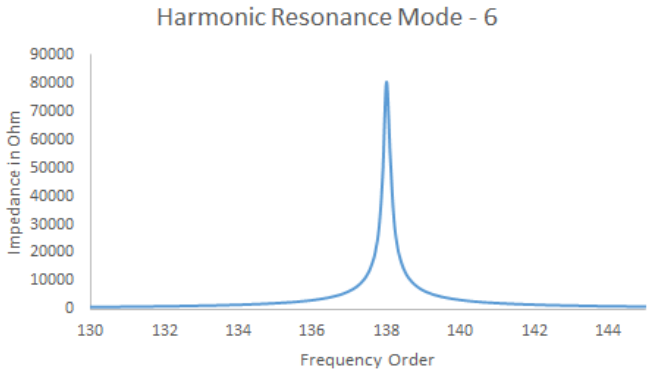


Figure 17: Harmonic Mode – 6, 138th Order

It is observed in figure – 14 and 15, Mode – 6 has now two resonance peak one is at 65.2nd order and second is at 138th order. The second peak is less affected by the relocation of PF capacitor but first resonance is not introduced, which was absent earlier. As the PF capacitor is now connected at 1.1 kV collector bus, it is affecting the network and it is the reason for reshaping of Mode – 6. Sensitivity of Modes is for Case – IV is given in Table – 5.

Table 5: Eigenvalue Sensitivity for Case-IV

Modal Sensitivity	Bus-1	Bus-2	Bus-3	Bus-4	Bus-5	Bus-6	Bus-7
Mode-3 (21.7)	0	0	0.3099	0.2915	0.3077	0.0877	0.0033
Mode-6 (65.2)	0.5	0.218	0	0	0.0014	0.0423	0.2375
Mode-6 (138)	0.5	0	0.0104	0	0	0.1661	0.3233

11. Filter Design

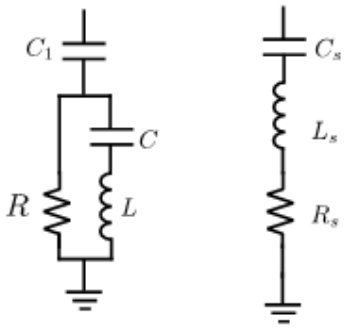


Figure 18: C-Filter and Equivalent Tuned Filter

$$L = \frac{m^2 - m\sqrt{m^2 - 1}}{h_0^2 - 1} \frac{2U_1^2}{Q_1 \omega_F} \quad (15)$$

The detail analysis of derivation of equation of 15 and 16 is given in reference [14]. The inclusion of C-Filter reduces the harmonic impedance to low level. For several value of m, the resonance point is checked for the stability and it is observed quite stable.

Table 6: Filter Parameter for Different Value of Cost Optimization Parameter

Sr. No	m	L (mH)	C(uF)
1	2	232.2	43.63
2	3	223	45.4
3	4	220	46
4	5	218.8	46.3
5	10	217.2	46.65

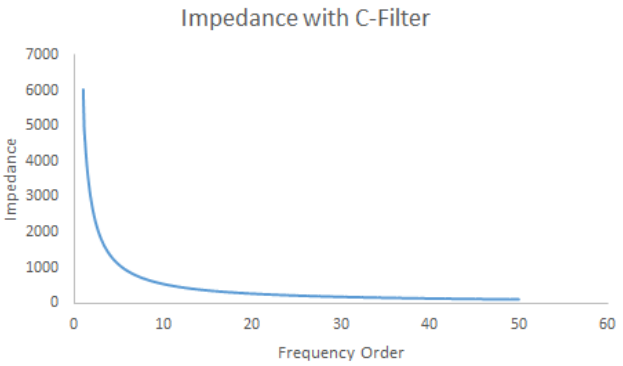


Figure 19: Impedance at 33 KV with R=450 Ohm

The impedance at 33 KV bus after inclusion of C-type filter is given in figure 15 for 450 Ohm of damping resistor. It reveals the important characteristics of C-type topology. The impedance is high around fundamental frequency and is reducing with frequency, so it offers lower impedance path for higher order harmonics and thus reduces the effect of resonance. The measurement data confirms the absence of such a high order harmonic in the inverter output, so the high impedance around 50th order will not affect the wind farm operation. The C-type topology mitigates the harmonic resonance effectively without any resurgence of new resonance point. The application of different topologies and their effectiveness is planned for the future work.

12. Conclusion

To recapitulate the core idea of this work, wind farms cause harmonic distortion in the power system. Also, the resonance point varies depends on the system parameters and system configuration. From the result of different cases, following points are concluded

1. As the wind power varies, the resonance point also varies, due to the change of power factor capacitor switching to maintain the power factor and also to support the system voltage.
2. The location of PF Capacitor will have significant effect on the resonance frequency
3. The resonance frequency reduces with increase in the size of Wind Farm.
4. The sensitivity of resonant point with respect to any component of power system is also demonstrated.
5. Sensitivity analysis helps us to plan and also design the harmonics filter to avoid critical conditions.

The optimized filter design using sensitivity analysis itself is a dedicated topic and future work on it is already planned. It is the reason why, it is not included in this work.

13. System Data

1. 1.1 kV Transformer: X=0.484 Ohm,
2. PF Capacitor: Xc=0.968 Ohm,
3. 1.1 kV Cable: R=0.01 Ohm, X=0.02 Ohm, C=0.36 uF,
4. 33 kV Transformer: X=8.712 Ohm,
5. 33 kV Cable: R=0.47 Ohm, X=1.6 Ohm, C=0.36 uF,
6. 132 kV Transformer: X =5.808 Ohm,
7. 132 kV Line: R=0.00080475 Ohm, X=0.01447 Ohm, R/X=18,
8. Grid: Fault Level=2500 MVA, R=0.3868 Ohm, X=0.9556 Ohm, R/X=18.

Conflict of Interest

We hereby declare that this work is based on the authors own idea and there is no conflict of interest.

References

[1] Sergio L. Varricchlo, Nelson Martins, Leonardo T. G. Ulma, "A Newton-Raphson Method Based On Elgenvalue Sensitivities to Improve Harmonic Voltage Performance", submitted for publication in the IEEE Transactions on Power Delivery.

[2] Sergio L. Varrichlo and Nelson Martins, "Filter Design Using a Newton-Raphson Method Based on Elgenvafue Sensitivity", IEEE Proceedings of the Surer Power Meeting, July 16-20, 2000, Seattle, Washington, USA.

[3] Kai Yang,Math H.J. Bollen, Fellow, IEEE, and Mats Wahlberg, "A comparison study of Harmonic Emission Measurements in Four Windparks", Paper to 2011 IEEE Power & Energy Society General Meeting. 978-1-4577-1002-5/11/$26. 00 ©2011 IEEE

[4] Kai Yang, *Student Member, IEEE,* Math H.J. Bollen, *Fellow, IEEE,* and Mats Wahlberg, "Comparison of Harmonic Emission at Two Nodes in a Wind Park", https://www.researchgate.net/publication/261055566, Conference Paper, DOI: 10.1109/ICHQP.2012.6381163

[5] Luis Sainz, Juan Jose Mesas, Remus Teodorescu, *Senior Member, IEEE*, and Pedro Rodriguez, *Member, IEEE, "*Deterministic and Stochastic Study of Wind Farm Harmonic Currents", IEEE Transactions On Energy Conversion, Vol. 25, No. 4, Pg 1071-1080, December 2010

[6] Paulo F. Ribeiro, *Fellow, IEEE,"*An Overview of Probabilistic Aspects of Harmonics:State of the Art and New Developments", IEEE Transaction on Power System, 2005.

[7] Fabio Medeiros, Dalton C. Brasil, Member, IEEE, Paulo F. Ribeiro, Fellow, IEEE, Cristiano A. G. Marques and Carlos A. Duque, Member, IEEE, "A New Approach for Harmonic Summation Using the Methodology of IEC 61400-21", DOI: 978-1-4244-7245-1/10/$26.00 ©2010 IEEE

[8] Morales, X. Robe and J.C. Maun, "Assessment of Wind Power Quality: Implementation of IEC61400-21 Procedures", RE&PQJ, Vol. 1, No.3, Pg 470 - 476, March 2005.

[9] Silas Y. Liu, Caio M. Pimenta, Heverton A. Pereira, Victor F. Mendes, Gabriel A. Mendonça And Selênio R. Silva, "Aggregated Dfig Wind Farm Harmonic Propagation Analysis", Anais do XIX Congresso Brasileiro de Automática, CBA 2012. ISBN: 978-85-8001-069-5

[10] Foroozan Ghassemi and Kah-Leong Koo, "Equivalent Network for Wind Farm Harmonic Assessments", IEEE Transactions On Power Delivery, Vol. 25, No. 3, Pg 1808-1815, July 2010.

[11] K.N Md Hasan, Kalle Rauma2, P. Rodriguez, J. Ignacio Candela, Raul S. Muñoz-Aguilar, Alvaro Luna, "An Overview of Harmonic Analysis and Resonances of Large Wind Power Plant", Pg 2467-2474,, 978-1-61284-972-0/11/$26.00 ©2011 IEEE

[12] Math H. J. Bollen And Kai Yang, "Harmonic aspects of wind power integration", J. Mod. Power Syst. Clean Energy (2013) 1(1):14–21, DOI 10.1007/s40565-013-0001-7

[13] Jignesh Patel and Satish K. Joshi, "Harmonic Stability of Grid Connected Renewable Energy Source- A Case Study", UPEC-2017, Greece, October 2017.

[14] Yao XIAO, Jie ZHAO and Shijie MAO "Theory of the Design of C-Type Filter", 11th Conference on Harmonic and Quality of Power (2014).

[15] S. L. Varricchio, N. Martins, L. T. G. Lima and S. Cameiro Jr."Studying Harmonic Problems Using a Descriptor System Approach", Proceedings of the IPST'99-International Conference on Power System Transients, Budapest, Hungary, June, 1999.

[16] P. M. Anderson, B. L. Agrawal and J. E. Van Ness, "Subsyncronous Resonance in Power System", IEEE Press, New York, USA, 1990.

[17] IEEE Harmonics Model and Simulation Task Force, "Test systems for harmonics modeling and simulation," *IEEE Trans. Power Del.*, vol. 14, no. 2, pp. 579–587, Apr. 1999.

[18] P. Kundur, Power System Stability and Control. New York: McGraw- Hill, 1994.

[19] G. T. Heydt, *Electric Power Quality*. West Lafayette, IN: Star in a Circle Publications, 1991

[20] R. C. Dugan, M. F. McGranaghan, and H. W. Beaty, *Electrical Power Systems Quality*. New York: McGraw-Hill, 1996.

[21] T. H. Ortmeyer and K. Zehar, "Distribution system harmonic design," *IEEE Trans. Power Delivery*, vol. 6, pp. 289–294, Jan. 1991.

[22] J. Martinon, P. Fauquembergue, and J. Lachaume, "A state variable approach to harmonic disturbances in distribution networks," in 7th Int. Conf. on Harmonics and Quality of Power, Las Vegas, NV, Oct. 16th–18th, 1996, pp. 293–299.

[23] Jignesh Patel, and Satish Joshi, "Harmonic Stability of Grid Connected Renewable Energy Source – A Case Study", UPEC 2017, Egypt.

[24] Verghese, G.C.; Perez-Arriaga, I.J.; Schweppe, F.C. "Selective modal analysis with applications to electric power systems—II. The dynamic stability problem," *IEEE Trans. Power App.Syst.*, vol. PAS-101, no. 9, pp. 3126–3134, Sep. 1982.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 08-20 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Simulation Strategy to Enhance Oxygen Uptake and Reaction Forces at Leg Joints and Vertebral Bodies During Ergometer Exercise Under Altered Gravity

Yoshihiko Tagawa*, Masayuki Omoto, Hiroo Matsuse, Naoto Shiba

School of Medicine, Orthopaedics, Kurume University, 830-0011, Japan

A R T I C L E I N F O

Article history:
Received: 05 June, 2018
Accepted: 22 June, 2018
Online: 02 July, 2018

Keywords:
Space exploration
Disuse atrophy
Astronaut
Elderly
Ergometer exercise
Hybrid training
Electrical stimulation

A B S T R A C T

Moderate exercise has been implemented in spaceflight programs to prevent muscle and joint damage under extreme circumstances such as microgravity. Although the physical effects of exercise have been investigated experimentally, advanced modeling and simulation techniques are powerful tools that could provide insight into the physical limitations, optimal conditions, and mechanisms that might lead to muscle and joint damage. Ergometers are safe for patients and elderly individuals because exercise intensity can be easily regulated. Moreover, ergometer exercise that removes the load imposed by body weight on leg joints would better reflect training conditions in locations such as the International Space Station or on the surfaces of the Moon and Mars, where the force of gravity is lower than that on the Earth. A hybrid training system (HTS) that combines the electrical stimulation of antagonist muscles and volitional contraction of agonist muscles is an effective training method. Co-contractile motions in the HTS can counteract the reduced effect of gravitational force on leg joints during ergometer exercise. The present study aimed to validate the ability of an ergometer exercise model to measure oxygen uptake and joint reaction forces by comparison with experimental results and create strategies for electrically stimulating leg joint muscles for modulation under altered gravity. The simulation results suggested that the HTS could easily control the magnitude of oxygen uptake and joint reaction forces. Optimal cycling conditions to achieve desirable values for oxygen uptake and joint reaction forces would help to maintain the health of astronauts and appropriate exercise programs could be implemented in constrained space facilities.

1. Introduction

Space exploration and aging induce physical deconditioning. Space exploration has been limited so far to a low Earth orbit and short visits to the Moon, but deconditioning will become more serious after lengthy exposure to microgravity and hypogravity in manned spaceflights to Mars and bases there and on the Moon. Effective and efficient countermeasures within limited spaces such as those in spaceships are indispensable for maintaining human health. This report extends the work that was originally presented at IEEE EMBC 2017 [1].

A method of creating countermeasures against microgravity and hypogravity must be established to safely generate effective outcomes. Ergometer exercise reduces the load imposed by gravitational force on leg joints and the metabolic cost is comparable to that of walking on a level surface [2]. This type of exercise is beneficial for persons with knee osteoarthritis and elderly individuals seeking to engage in cardiovascular training. However, such exercise is ineffective for maintaining musculoskeletal function during spaceflight because the mechanical load provided by aerobic exercise is too low to prevent muscle atrophy at 0 g [3]. Preventing bone loss requires a compressive load along the bone axis similar to the repetitive forces at work during routine activities on the Earth. Astronauts are instructed to execute resistance and endurance exercises for about two hours and day for six days a week in a combination of countermeasures against the effects of microgravity. The conflicting requirements of safe aerobic training with less intensive loads on leg joints must be satisfied under lengthy exposure to microgravity. Creating appropriate experimental approaches on the Earth and under microgravity and hypogravity is very challenging.

*Yoshihiko Tagawa, Email: tagawa.yoshihiko877@mail.kyutech.jp

Although the effects of exercise on the human body have been investigated using mostly experimental approaches, advanced modeling and simulation techniques could be powerful tools for gaining insight into issues such as human physical limitations, forces at joints and forces generated by contracting muscles and the optimal conditions for training or exercise. Such techniques will help to quantitatively and qualitatively predict the effects of exercise under conditions of altered gravity. The present study used ergometer exercise models with the AnyBody Modeling System (AMS) ver. 6.0.6 (AnyBody Technology A/S, Aalborg, Denmark).

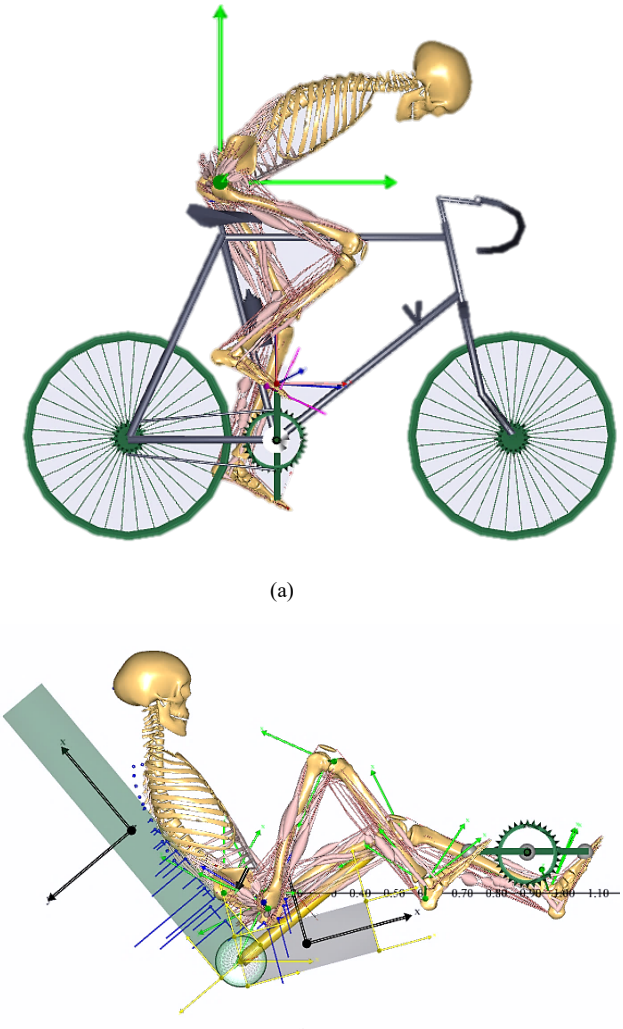
Oxygen uptake and loading on leg joints during ergometer exercise have only been examined experimentally. The present study aimed to validate an ergometer exercise model over a wide range of conditions to estimate oxygen uptake and reaction forces at leg joints and vertebral bodies by comparing the outcomes with experimental results, to develop a strategy for electrically stimulating antagonist muscles of joint motion to modulate oxygen uptake and joint loading under altered gravity, and to determine optimal conditions to satisfy conflicting requirements. A hybrid training system (HTS) that combines the electrical stimulation and volitional contraction of antagonist and agonist muscles, respectively, is an effective type of training [4] even for upper limbs under microgravity [5]. Co-contractile motions in the HTS can counteract the reduced effect of the gravitational force on leg joints during ergometer exercise.

2. Simulation Models for Ergometer Exercise

The musculoskeletal model in the AMS replicates a 50th percentile European male with a mass of 75 kg and a height of 1.76 m, respectively [6]. We modified the muscle activity function and muscle strength of "BikeModel" [7] in the AnyBody Managed Model Repository (AMMR ver.1.6.3) to construct a conventional ergometer (CER) exercise model with the HTS and reduced muscle strength. A seated model was constructed from the "StandingModel" based on the "BikeModel" and "LegPressMachine" in the AMMR ver.1.6.3 for recumbent ergometer (RER) exercise. Figure 1 shows both models that can simulate cycling with electrical stimulation of muscles around the leg joints based on the HTS concept and reduced muscle strength to represent weakened muscles. External forces and moments were applied to the thorax to maintain upper body posture while cycling on the CER model, even though the model lacks arms. In the RER model, a constant force of 56 N was applied to both anterior superior iliac spines to fix the pelvis on the seat of the ergometer. Friction forces (static friction coefficient of 0.6) between the surfaces of the backrest and the thoracic region, and between the gluteus region and the seat were also assumed instead of external forces and moments at the thorax. Thus, a person using the RER could exercise even at a gravitational acceleration < 1 g (9.81 m/s^2) on the Earth. The relative positions of the hip joint and crank axis of the RER were determined to mimic that of the CER. The ankle plantar angle on the RER was obtained experimentally from our recumbent ergometer cycling with a position mimicking that of CER exercise at a pedal rate of 60 rpm and a mechanical load of 60 W. The leg joint angles of both models became similar (Figure 2). A crank torque on the pedal of the RER has a phase shift of quarter cycle of a crank torque on that of the CER.

The amount of oxygen uptake during exercise on the models can be derived from the caloric equivalent of oxygen [8] required for muscle power. The simulation results included oxygen uptake of 3.5 ml/kg/min at rest only when compared with the experimental results that were plotted with reference to regression equations derived from the literature.

Recruitment patterns in a redundant muscle system such as the human body are produced based on optimal methods to minimize an objective function and they consist of the ratio of muscle force to maximal muscle force; that is, muscle activity. The AnyBody system offers min/max criteria for muscle recruitment [9]: the largest endurance for a task or identical activation of all muscles contributes positively to balancing external force, thus minimizing maximal muscle activity as much as possible when muscles work together. This would be a physiological criterion for minimal fatigue [10]. We combined the quadratic term of muscle activity with the objective function. A Hill-type muscle model [11] was adopted for the leg of the exercise model.



(a)

(b)

Figure 1. Simulation models for ergometer cycling with hybrid training system (HTS) and reduced muscle strength. (a) Conventional ergometer (CER) from "BikeModel" in the AnyBody Managed Model Repository (AMMR ver.1.6.3). (b) Recumbent ergometer (RER) constructed from CER with backrest angle of 130° and seat tilt of 15°.

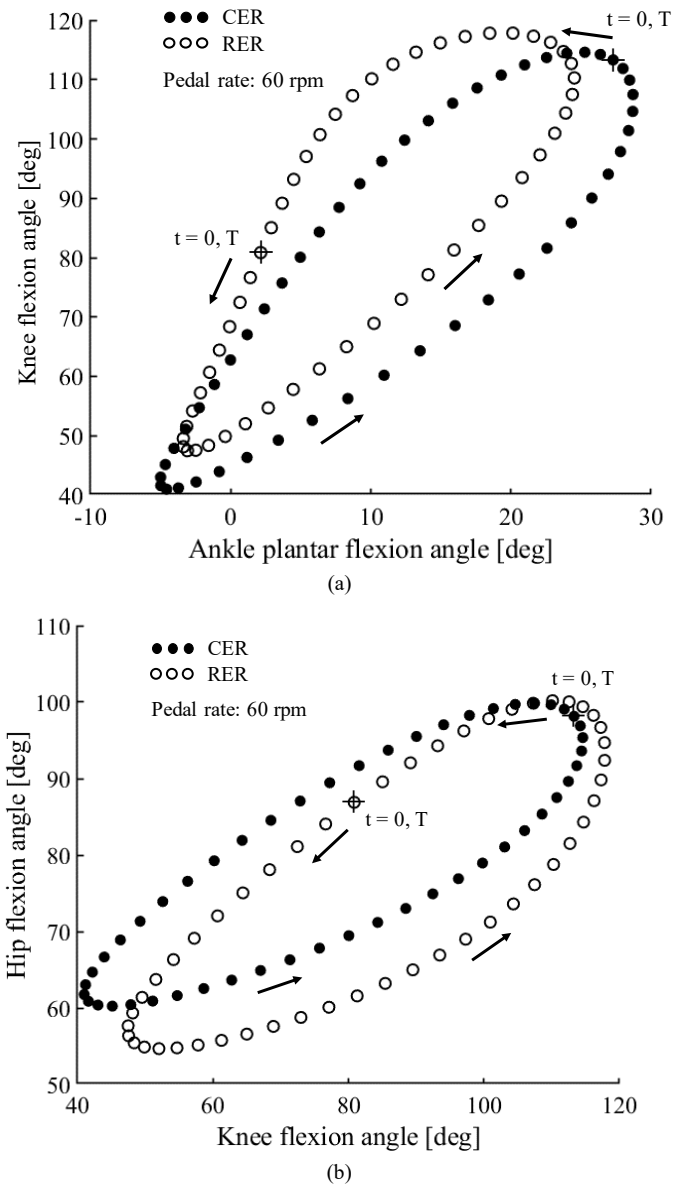


Figure 2. Comparisons of leg joint angles between conventional ergometer (CER) and recumbent ergometer (RER) exercise at pedal rate of 60 rpm. (a) Ankle plantar flexion vs. knee flexion. (b) Knee flexion vs. hip flexion. T, pedaling duration. At time 0, the right foot is located at highest point.

3. Validation of Ergometer Exercise Models

Volitional CER (VCER) and volitional RER (VRER) exercise models were validated based on comparisons with experimental results on the Earth. The saddle height and position were set to 0.78 m superior and 0.17 m posterior from the crank axis on the VCER. The center of the hip joint was positioned 0.77 m posterior and 0.15 m inferior from the crank axis on the VRER. The length of both cranks was 0.17 m. The effect of the HTS combined with the CER (HCER) was also investigated.

3.1. VCER Exercise at 1 g

The pedal rate and mechanical load of the VCER exercise model were varied with expanded experimental ranges. Figure 3 and 4 show oxygen uptake (ml/kg/min) at a pedal rate of 60 rpm under variable mechanical loads, and at defined mechanical loads of 81.7 and 0 W under different pedal rates, respectively. The simulation results shown in Figure 3 closely agreed with the experimental results [12] and increased linearly as mechanical load increased. The simulation results shown in Figure 4 were also similar to the experimental results [13,14] and increased

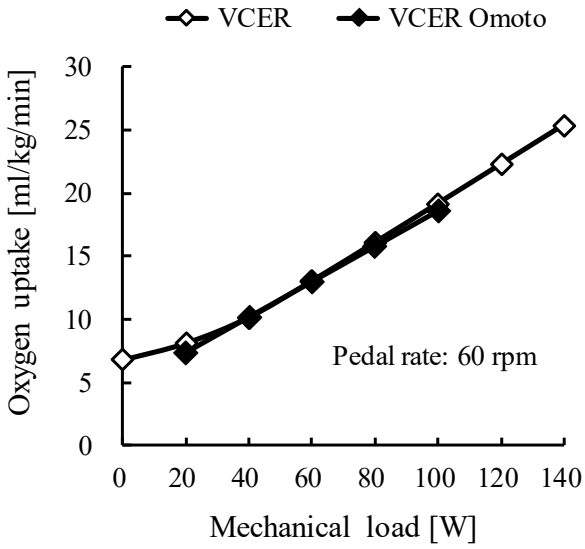


Figure 3. Oxygen uptake during volitional conventional ergometer (VCER) exercise at a pedal rate of 60 rpm under various mechanical loads. Unfilled and filled symbols indicate simulated and experimental results, respectively. Omoto, oxygen uptake data from [12].

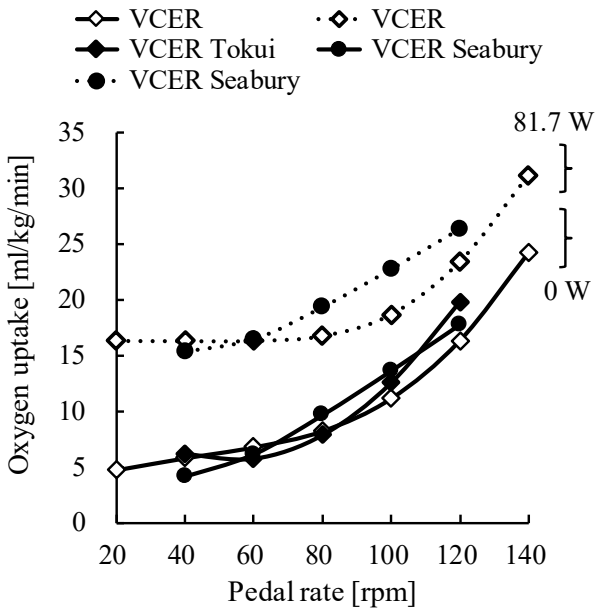


Figure 4. Oxygen uptake during volitional conventional ergometer (VCER) exercise under various mechanical loads (0 and 81.7 W) and pedal rates. Unfilled and filled symbols indicate simulated and experimental results, respectively. Tokui and Seabury, oxygen uptake data from [13] and [14].

nonlinearly, showing quadratic curves as the pedal rate increased. Oxygen uptake will consistently increase beyond the upper limit of the experimental range.

The maximum magnitude of resultant forces at the knee joint during cycling has been measured using instrumented knee implants in elderly individuals who were free of pain and physically active after total knee replacement (TKR) [2]. We

assumed 25% less muscle strength in the leg of the VCER exercise model with the muscle recruitment criterion of combined min/max and quadratic forms of muscle activity for comparisons with measured results. The simulated magnitude of the reaction force was reduced by approximately 10% of that using 100% muscle strength in the model. Moreover, the saddle height was set to 0.815 m, which mimicked the experimental conditions [2]. The simulated knee joint force closely agreed with the experimental findings of elderly individuals who engaged in ergometer exercise with an increasing mechanical load (Figure 5). The slow pedal rate resulted in a larger knee joint force. The increase in knee joint reaction force might be consistent with that above the experimental range.

The similar findings of these comparisons supported the validity and reliability of the ergometer exercise model except when values were beyond the lower limit of the experimental range. Little is known about ergometer exercise experiments at lower mechanical loads and pedal rates.

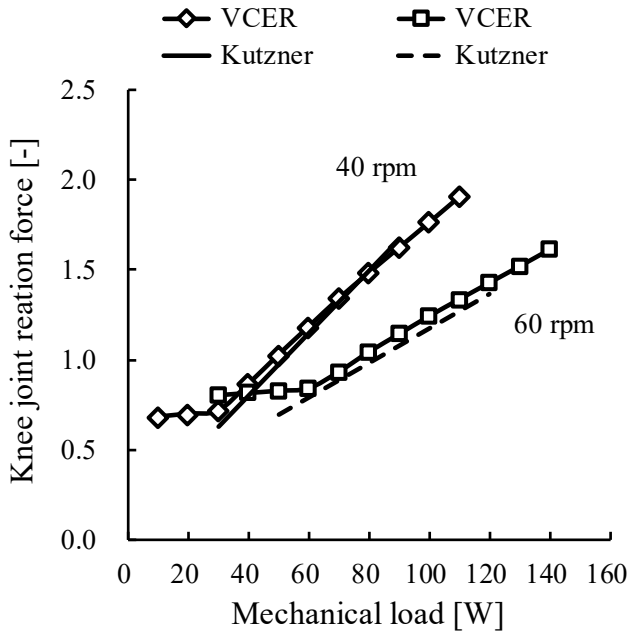


Figure 5. Comparison of experimental and simulated results of knee joint reaction force during volitional conventional ergometer (VCER) exercise under various mechanical loads and pedal rates of 40 and 60 rpm. Reaction force is maximum magnitude of force normalized by body weight. Saddle height in simulation was set to 0.815 m to mimic experimental condition. Kutzner, joint reaction force data from [2].

3.2. VRER Exercise at 1 g

Although both models had similar angles of leg joints, the upper body in the CER was dropped forwards and the body in the RER was inclined backwards (recumbent). Mean oxygen uptake at a pedal rate of 75 rpm in both positions was essentially identical at the same mechanical load [15]. Consumption tended to increase at higher loads in the recumbent position [15]. Simulated oxygen uptake in the VCER and VRER models was compared from the perspective of different mechanical loads and pedal rates (Figure 6.)

The difference in oxygen uptake between the two types of exercise rapidly increased as pedal rates increased, with oxygen uptake being higher for VRER, than VCER exercise. Further investigation is needed to validate the effect of RER exercise on the body.

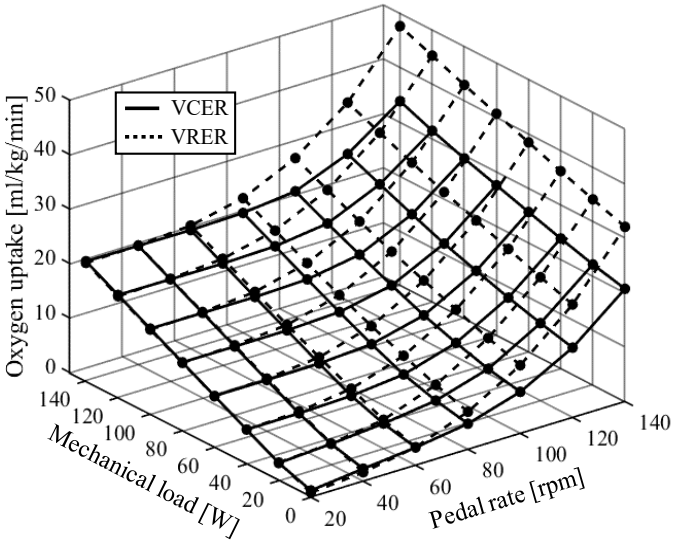


Figure 6. Comparison of simulated oxygen uptake during volitional conventional (VCER) and recumbent (VRER) ergometer exercise under different mechanical loads and pedal rates. Solid and dotted lines indicate VCER and VRER exercise, respectively. More oxygen was consumed during VRER than VCER exercise. Both results increased linearly according to increasing mechanical load and nonlinearly according to increasing pedal rate.

3.3. HCER Exercise at 1 g

The HTS was applied to the leg joints during ergometer exercise using the same leg motion as that used for volitional cycling. Stimulation intensities on the quadriceps femoris and hamstrings of the simulation model were determined at 1% and 5% of the maximum muscle activity, respectively, using the lower boundary of the muscle activity function in AnyBody. Stimulation that can fully activate assigned antagonist muscles at a specific intensity is ideal, unlike the actual situation during surface electrical stimulation (SES) [16-19]. The results of the simulation with low-intensity HTS closely agreed with the experimental findings of high-intensity SES (Figure 7). Additional effects of stimulation at intensities of 10%, 2% and 10% to the gastrocnemius, soleus, and tibialis anterior, respectively, were simulated to induce more oxygen uptake. At a pedal rate of 60 rpm (1-s cycle time), the antagonist muscles of the right knee during extension and flexion at the knee joint were stimulated for specific durations when knee angular velocities were > 2.0 and < –2.0 rad/s, respectively. These durations corresponded to 34% of the cycle time. The threshold of angular velocity of the plantar flexion angle of the right ankle was set to 1.0 rad/s, the corresponding duration of which was 26% of the cycle duration. The onset of stimulation on the left knee and ankle joints shifted half a cycle. The stimulus durations at all pedal rates were kept constant by adjusting the angular velocity setpoint according to the rate.

The increased oxygen uptake in the HCER was caused mainly by muscle co-contraction at the knee and ankle joints. The HCER linearly increased oxygen uptake, and the additional stimulation of muscles at the ankle joint similarly increased oxygen uptake.

The increase in oxygen uptake with the HTS shows that the HCER can achieve the same exercise intensity as the VCER at higher mechanical loads.

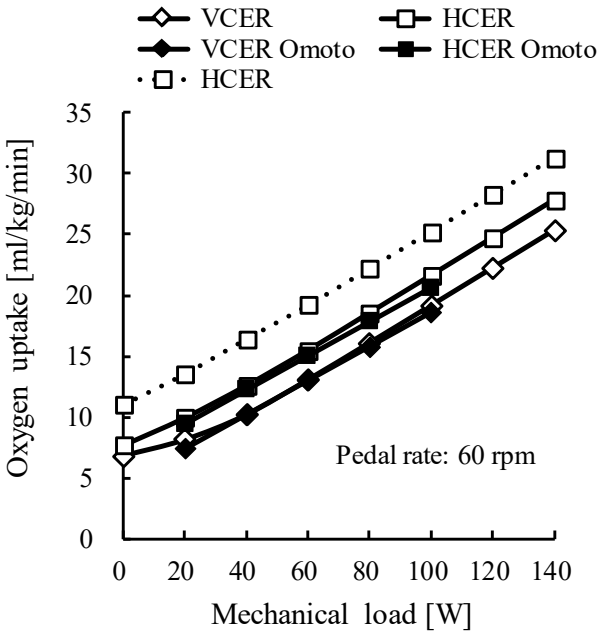


Figure 7. Comparison of simulated and experimental results of oxygen uptake during volitional conventional ergometer (VCER) exercise and hybrid training system (HTS) combined with conventional ergometer (HCER) exercise under different mechanical loads at pedal rate of 60 rpm. Solid line, HTS acting on knee joint; dotted line, HTS acting on ankle and knee joints. Unfilled and filled symbols indicate simulated and experimental results, respectively. Omoto, oxygen uptake data from [12].

4. Prediction Under Altered Gravity

The simulation and experimental results at the acceleration of gravity on the Earth closely agreed. The effects of different accelerations of gravity, such as microgravity in spaceships and hypogravity on the surfaces of Mars and the Moon, on the human body must be considered for safe space exploration, maintaining health under altered gravity, and returning smoothly to activities of daily life (ADL) on the Earth. Good agreement will support various predictions under altered gravity.

4.1. Oxygen Uptake During VCER and VRER Exercise Under Altered Gravity

The amounts of oxygen uptake at 0 g under mechanical loads of 60 and 120 W were essentially identical for each exercise. As gravity increased, oxygen consumption assumed a concave form during VCER exercise and a moderately concave form at both loads during VRER exercise. Although oxygen uptake varied, but the difference remained the same independently of gravity [20]. The lower oxygen uptake for the VCER and VRER ranged from 0.7 to 1.2 g and was dependent to some degree on mechanical load (Figure 8). The range of moderate gravity seemed to play assistive and resistive roles in the opposite leg for efficient oxygen consumption. Oxygen uptake during VCER exercise under varying gravitational acceleration became more concave than that while on the VRER. This might have been due to the different structure between the VCER and VRER; namely, the coordination of muscle activities of both legs is enhanced to maintain the pelvis position and steady cycling on the VCER under hypo- and hypergravity, respectively. In contrast, the backrest of the RER seat supports a reaction against the forward pushing force on the pedal of the VRER and restrained activities of lower limb muscles at hypergravity.

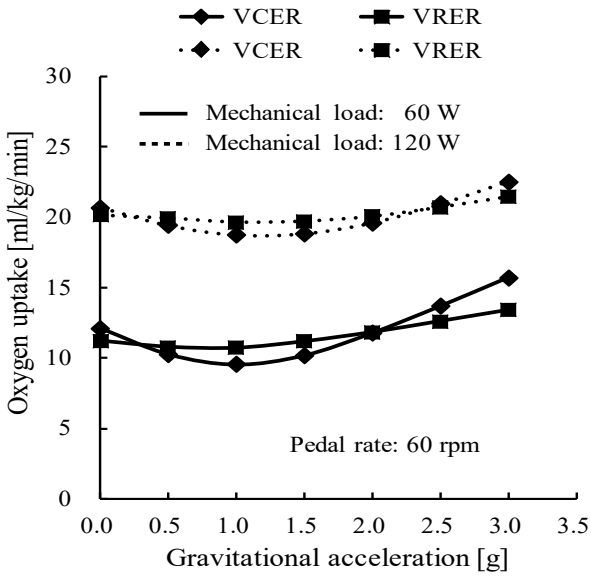


Figure 8. Simulated oxygen uptake under mechanical loads of 60 and 120 W and altered gravitational acceleration during volitional conventional ergometer (VCER) exercise and volitional recumbent ergometer (VRER) exercise. g, acceleration of gravity on the Earth; 1 g = 9.81 m/s^2. Pedal rate, 60 rpm.

4.2. Oxygen Uptake and Joint Reaction Force During VCER and VRER Exercise at 0 g

Oxygen uptake and the maximum magnitude of leg joint reaction forces that were normalized by body weight in the model, increased linearly with increasing mechanical load during VCER (Figure 9) and VRER exercise (Figure 10). Oxygen uptake and joint reaction forces at the knee and hip joints during VCER exercise were larger overall at 0 g (dotted lines) compared with those at 1 g (solid lines). Knee reaction force during VRER exercise was larger overall at 0 g than at 1 g, and oxygen uptake slightly increased with moderate variations under altered gravity.

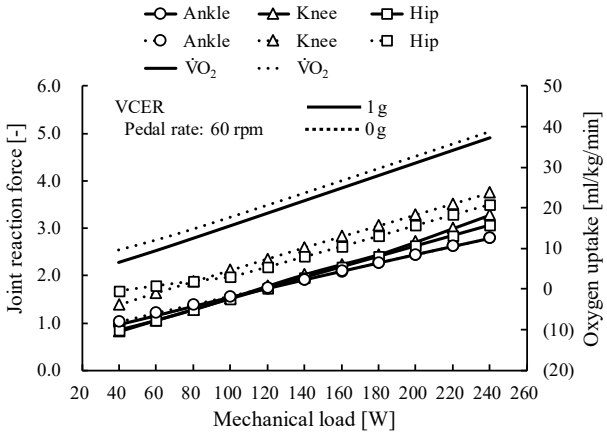


Figure 9. Comparison of joint reaction forces and oxygen uptake between 1 g (solid line) and 0 g (dotted line) during volitional conventional ergometer (VCER) exercise at pedal rate of 60 rpm. Joint reaction force is maximum magnitude of force normalized by body weight. $\dot{V}O_2$, oxygen uptake.

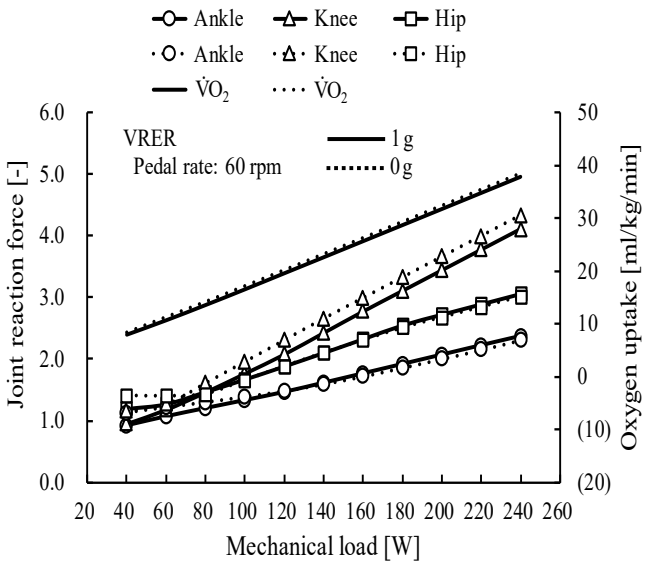


Figure 10. Comparison of joint reaction forces and oxygen uptake between 1 g (solid line) and 0 g (dotted line) during volitional recumbent ergometer (VRER) exercise at pedal rate of 60 rpm. Joint reaction force is maximum magnitude of force normalized by body weight. $\dot{V}O_2$, oxygen uptake.

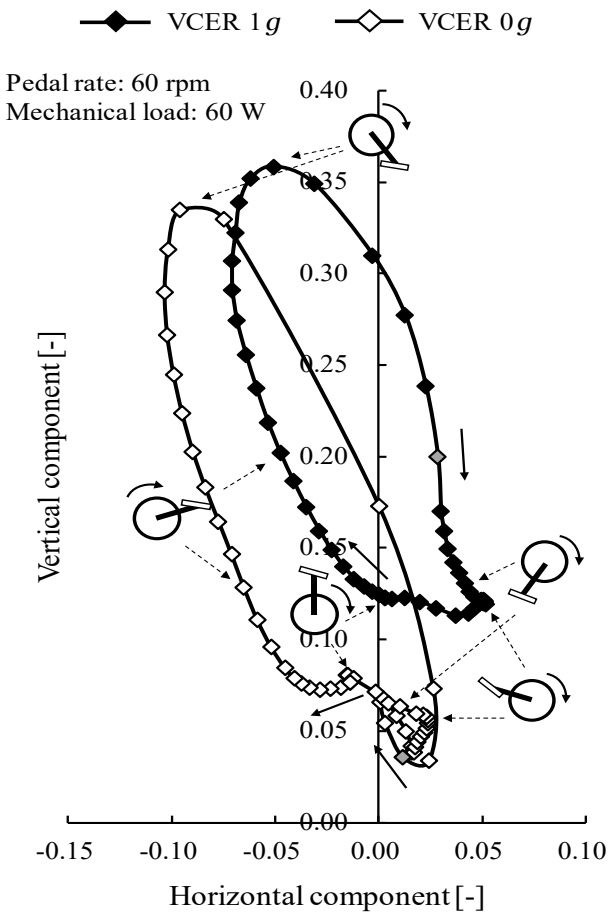


Figure 11. Comparison of pedal reaction forces between 1 g (filled symbols) and 0 g (unfilled symbols) during volitional conventional ergometer (VCER) exercise at a pedal rate of 60 rpm and mechanical load of 60 W. Pedal reaction force is shown in horizontal and vertical components, normalized by body weight. Gray symbol, lowest point of pedal.

4.3. Compensatory Actions During VCER Exercise at 0 g

The vertical component of the pedal reaction force became reduced and the horizontal component increased at 0 g (Figure 11), and the resultant pedal reaction force was lower than that at 1 g. The simulated force perpendicular to the crank (effective force) in the sagittal plane during VCER exercise was lower at 0 g with a pedal rate of 60 rpm and a mechanical load of 60 W than that at 1 g (Figure 12).

Muscles in the leg are activated to compensate for gravitational loss (Figure 13). Muscle force and joint reaction forces increase at 0 g and the increase in oxygen uptake was enhanced. The compensatory actions were smaller in the VRER than in the VCER, but the trends were similar.

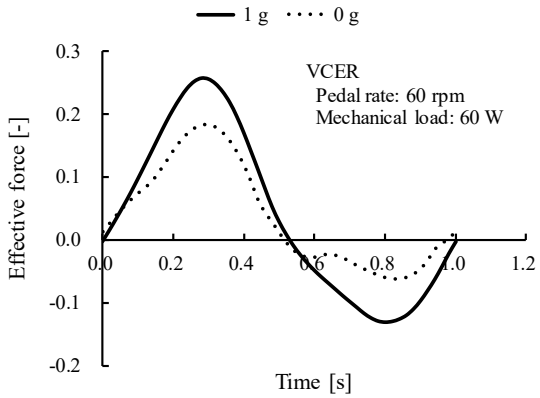


Figure 12. Comparison of effective force between 1 g (solid line) and 0 g (dotted line) during volitional recumbent ergometer (VCER) exercise at pedal rate of 60 rpm and mechanical load of 60 W. Effective force perpendicular to crank, normalized by body weight.

4.4. Effect of HTS on Ergometer Exercise at 0 g

Figure 14 shows the simulation results of oxygen uptake and knee joint reaction force during HCER exercise at 0 g with a mechanical load and two pedal rates corresponding to the experiment shown in Figure 5 at 1 g. The stimulation intensity was 10% and 20% of the maximum muscle activity of the quadriceps femoris and hamstrings, respectively. The oxygen uptake during VCER at 0 g varied linearly and increased at a pedal rate of 60 rpm.

Oxygen uptake under HTS will vary linearly with the same increment at mechanical loads that exceed the upper limit of the experimental range (similar to what is shown in Figure 9 for the VCER). The maximum magnitude of knee joint reaction forces during HCER exercise was over 3-fold body weight, which was much larger than the experimental and simulation results within the range of the mechanical load shown in Figure 15.

We confirmed the effects of the HTS on the reaction forces at leg joints over the mechanical load range of the experiment (Figure 15). The quadriceps femoris and hamstrings were stimulated at the same intensity as that shown in Figure 14. The maximum magnitude of resultant reaction forces induced at the knee and hip joints were ≥ 3-fold body weight and equivalent at least to the maximum at a mechanical load of 200 W during VCER exercise at 0 g (Figure 9). The maximum reaction forces at the leg joints at 40 rpm were essentially larger than those at 60 rpm.

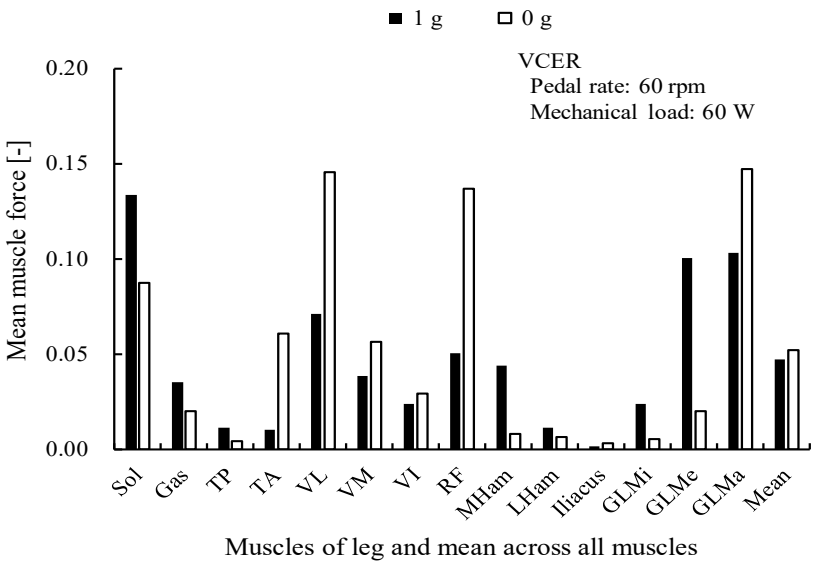


Figure 13. Comparison of mean forces of leg muscles between 1 g (filled bars) and 0 g (unfilled bars) during volitional conventional ergometer (VCER) exercise at pedal rate of 60 rpm and mechanical load of 60 W. Force is averaged during one cycle, normalized by body weight. Mean, average of mean muscle force of leg at 0 g and became larger than that at 1 g. Abbreviations: Gas, gastrocnemius; GLMa, gluteus medius; GLMe, gluteus medius; GLMi, gluteus minimus; LHam, lateral hamstrings; MHam, medial hamstrings; RF, rectus femoris; Sol, soleus; TA, tibialis anterior; TP, tibialis posterior; VI, vastus intermedius; VL, vastus lateralis; VM, vastus medialis.

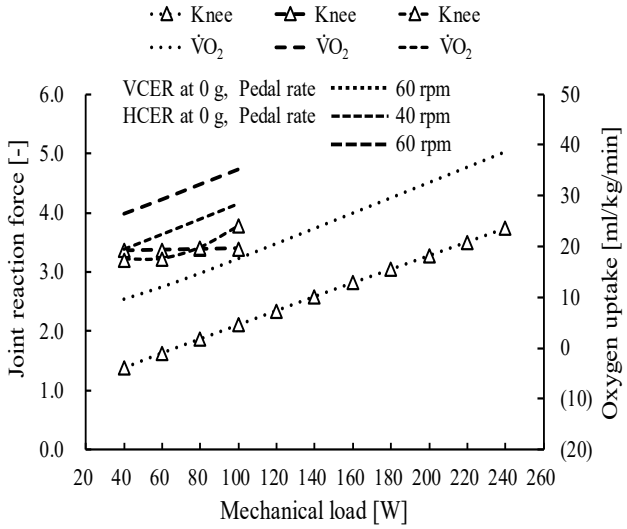


Figure 14. Comparison of knee joint reaction force and oxygen uptake at 0 g during volitional conventional ergometer (VCER) exercise and hybrid training system (HTS) combined with conventional ergometer (HCER) exercise under various mechanical loads and pedal rates of 40 and 60 rpm. Joint reaction force is maximum magnitude of force normalized by body weight. $\dot{V}O_2$, oxygen uptake.

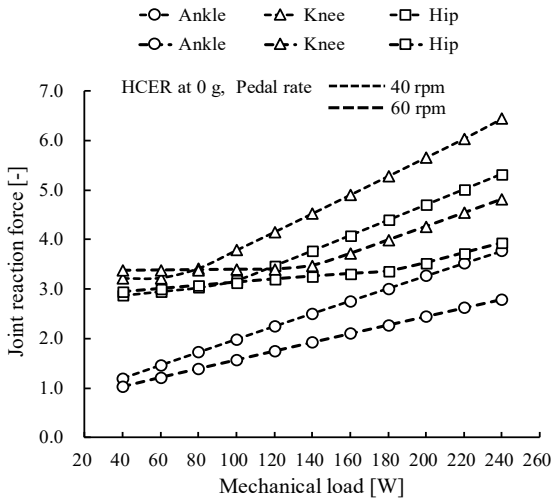


Figure 15. Comparison of joint reaction forces at 0 g during hybrid training system (HTS) combined with conventional ergometer (HCER) exercise under different mechanical loads and pedal rates of 40 and 60 rpm. Joint reaction force is maximum magnitude of force normalized by body weight.

The HTS increased the amount of oxygen uptake and knee joint reaction forces. The simulated results of oxygen uptake and knee joint reaction force during VCER and HCER exercise at 0 g were comprehensively compared under various mechanical loads and pedal rates (Figures 16 and 17, respectively). The amount of variation in the incremental pedal rate was contradictory. The quadriceps femoris and hamstrings of the HCER model were stimulated at intensities of 10% and 20% of the maximum muscle activity, respectively.

The optimal cycling conditions of mechanical load and pedal rate required to achieve a specific intensity and knee joint reaction force can be determined by minimizing the objective function, f (1). Table 1 shows the calculated optimal values for VCER and HCER exercise at 0 g. The reaction force was normalized by the body weight of the model.

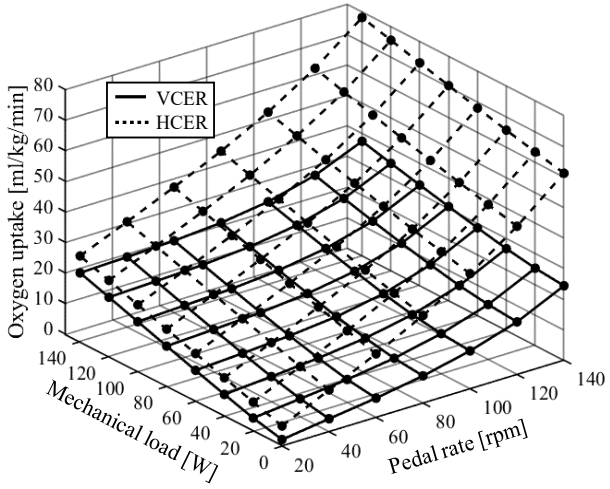


Figure 16. Simulated results of oxygen uptake at 0 g during volitional conventional ergometer (VCER) exercise and the hybrid training system (HTS) combined with conventional ergometer (HCER) exercise under various mechanical loads and pedal rates. Solid and dotted lines, VCER and HCER exercise, respectively. More oxygen was consumed during HCER than VCER exercise.

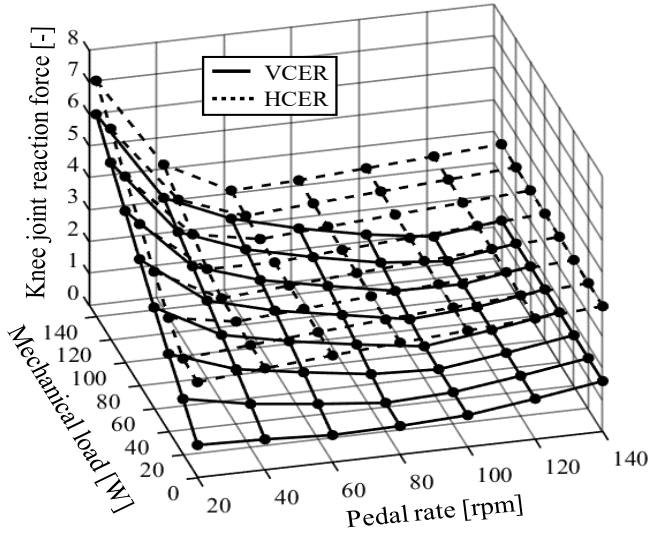


Figure 17. Simulated results of knee joint reaction at 0 g during volitional conventional ergometer (VCER) exercise and the hybrid training system (HTS) combined with conventional ergometer (HCER) exercise under various mechanical loads and pedal rates. Solid and dotted lines, VCER and HCER exercise, respectively. Knee joint reaction force is greater at lower pedal rate and greater mechanical load in both exercises. Joint reaction force is maximum magnitude of force normalized by body weight.

$$f = w_1\left(\dot{V}O_{2,m} - \dot{V}O_{2,d}\right)^2 + w_2\left(F_{Kj,m} - F_{Kj,d}\right)^2 \quad (1)$$

$\dot{V}O_{2,*}$: oxygen uptake, $F_{Kj,*}$: reaction force at knee joint normalized by the body weight of the model, *: m, model; d, desired value, w_1, w_2: weight coefficients

Table 1. Optimal cycling conditions for VCER and HCER at 0 *g*

Model	Optimal cycling conditions		Desired / realized values	
	Mechanical load [W]	***Pedal rate*** [rpm]	***Oxygen uptake*** [ml/kg/min]	***Reaction force at knee joint*** [-]
VCER	120	42.5	20/20	3.0/3.0
HCER	54.9	35.2	20/20	3.0/3.2
VCER	182	61.7	30/30	3.0/3.0
HCER	92.7	48.0	30/30	3.0/3.3
VCER	243	77.7	40/40	3.0/3.0
HCER	133	59.6	40/40	3.0/3.4

HCER, hybrid training system combined with conventional ergometer; VCER, volitional conventional ergometer

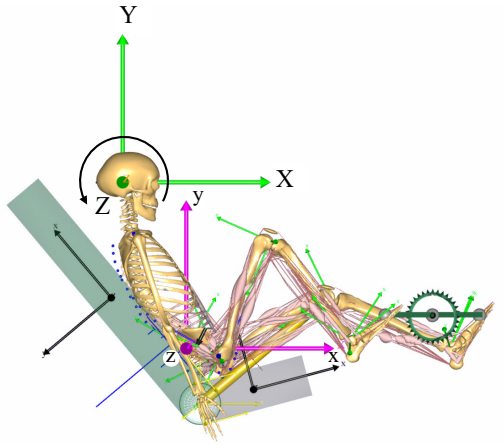


Figure 18. Human-ergometer system. System rotates about the Z axis of O-XYZ. O-XYZ, inertial frame; o-xyz, frame of human-ergometer system with origin at pelvis.

Oxygen uptake and joint reaction forces were simulated by altering the rotational velocity of the RER model at 0 g, a pedal rate of 60 rpm and a mechanical load of 60 W (Figure 18). We assumed that the upper limb in this model reflected the force of artificial gravity at a fixed shoulder joint. The rotational axis Z was set to pass close to both ears. The RER rotated at a constant angular velocity at increments of 90 deg/s from 0 to 450 deg/s. The centrifugal acceleration of the origin of the human ergometer frame at each velocity (o-xyz), was 0, 0.16, 0.65, 1.45, 2.59, and 4.04 g. The HTS was introduced at the knee and ankle joints to increase oxygen uptake and joint reaction forces at the ankle, knee, and hip joints.

Figure 19 shows the maximum magnitude of resultant forces between adjacent vertebral bodies located below the 12th thoracic vertebra. The resultant forces were greater during the HTS combined with the RER (HRER) than VRER exercise at angular velocities of 0 and 90 deg/s. Both results increased quadratically with increases in the angular velocity of the human ergometer system, in which the respective artificial gravity of which was quadratically proportional to the angular velocity. The resultant forces between adjacent vertebral bodies remarkably increased with increasing artificial gravity to > 360 deg/s that generated a centrifugal acceleration of 2.59 g, and were close to the resultant forces that are generated when walking on a level surface [21].

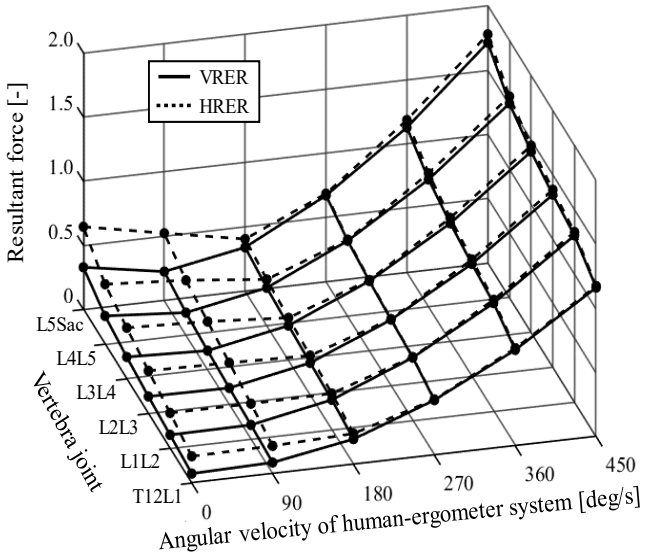


Figure 19. Comparison of peak resultant forces between adjacent vertebral bodies during volitional recumbent ergometer (VRER) exercise and the hybrid training system combined with recumbent ergometer (HRER) exercise at 0 g, a pedal rate of 60 rpm and mechanical load of 60 W. Solid and dotted lines, VRER and HRER exercise, respectively. Reaction force is the maximum magnitude of the force, normalized by body weight. Stimulation intensity, 10%, 20%, 20%, 4% and 20% at quadriceps, hamstrings, gastrocnemius, soleus and tibialis anterior, respectively. L5Sac, between 5th lumbar spine and sacrum; LiLi+1, between ith and (i+1)th lumber vertebrae (i = 2,3,4); T12L1, between 12th thoracic and 1st lumbar vertebrae.

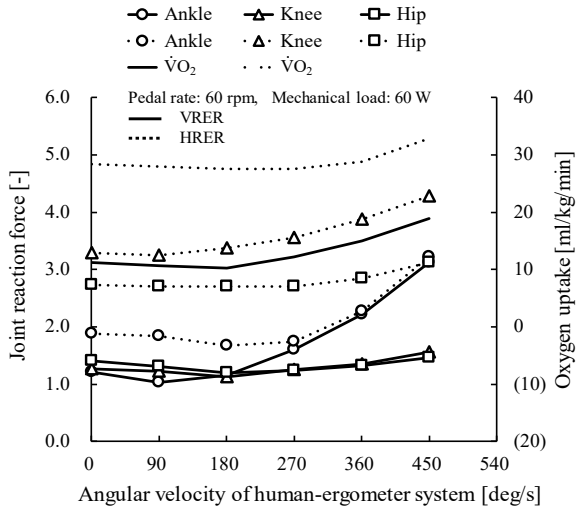


Figure 20. Comparison of peak joint reaction forces and oxygen uptake during volitional recumbent ergometer (VRER) exercise and the hybrid training system combined with recumbent ergometer (HRER) exercise at 0 g, pedal rate of 60 rpm and mechanical load of 60 W. Solid and dotted lines, VRER and HRER exercise, respectively. Mean, average of mean leg muscle forces. Stimulation intensity, 10%, 20%, 20%, 4% and 20% at quadriceps, hamstrings, gastrocnemius, soleus and tibialis anterior, respectively. Joint reaction force is maximum magnitude of force, normalized by body weight.

5. Discussion

Human in an environment with microgravity will experience disuse atrophy, resulting in the loss of mass and strength in muscles and bones. Exercise under such extreme circumstances has been examined and moderate exercise intensity has been implemented in spaceflight programs. Conventional ergometer cycling is a safe type of exercise that is easily implemented, because the exercise intensity can be controlled. Advanced modeling and simulation techniques are powerful tools that can provide insight into the physical limitations, optimal conditions, and mechanisms that might lead to muscle and joint damage instead of experimental approaches under extreme environments. Few studies have applied mathematical modeling and experimental approaches to determine joint reaction forces while walking [22] and oxygen uptake during ergometer exercise [20]

The cycling exercise model was constructed by modifying a commercially available bicycle model to simulate oxygen uptake during HTS cycling at 1 g and 0 g. The simulated oxygen uptake during VCER and HCER exercise at 1 g closely agreed with the experimental results described in [12] (Figure 7). Mean oxygen uptake at a pedal rate of 75 rpm in dropped forward and recumbent positions is essentially identical at the same mechanical load, but slightly increased at higher loads in the recumbent position [15]. Since we confirmed the same tendency in the VRER (Figure 6), several features of RER exercise predicted by simulation seemed acceptable. Oxygen uptake in hypergravity was lower during VRER than VCER exercise. This might have been due to the coordination of muscle activities in both legs being enhanced to maintain the steady crank rotation during VCER exercise, whereas reaction force generated from the back rest contributes to adjust the moment at the leg joints in the VRER. Consequently, the mean muscle forces of the legs in the VRER at hypergravity of 3 g became moderate and averaged far less than those in the VCER (Figure 21). Large muscle forces of the soleus (one-joint muscle across the ankle) were generated to balance the large pedal reaction forces at high gravitational and centrifugal acceleration.

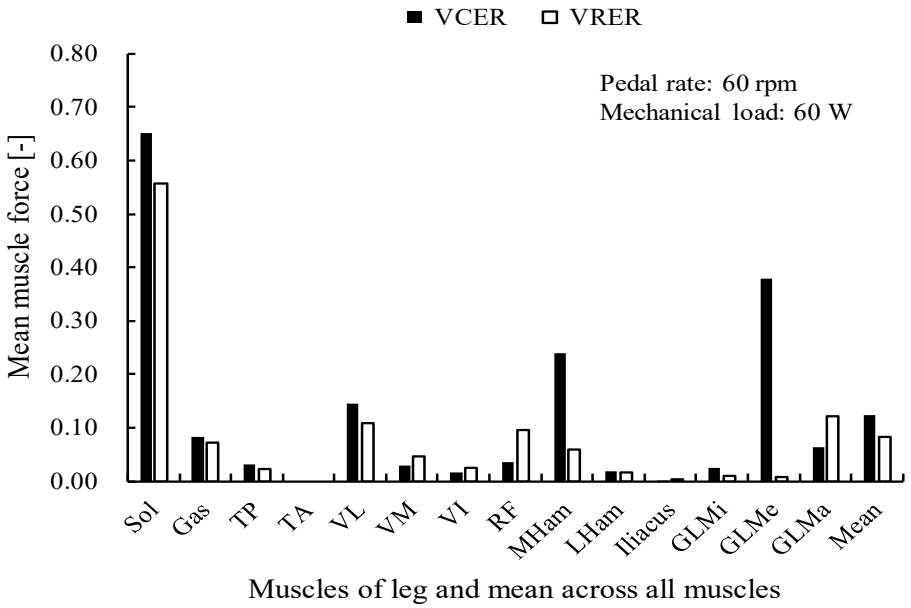


Figure 21. Comparison of mean leg muscle forces during volitional conventional (VCER, filled bars) and volitional recumbent (VRER, unfilled bars) ergometer exercise at 3 g, a pedal rate of 60 rpm and mechanical load of 60 W. Force was averaged during one cycle, normalized by body weight. Abbreviations: Gas, gastrocnemius; GLMa, gluteus medius; GLMe, gluteus medius; GLMi, gluteus minimus; LHam, lateral hamstrings; MHam, medial hamstrings; RF, rectus femoris; Sol, soleus; TA, tibialis anterior; TP, tibialis posterior; VI, vastus intermedius; VL, vastus lateralis; VM, vastus medialis.

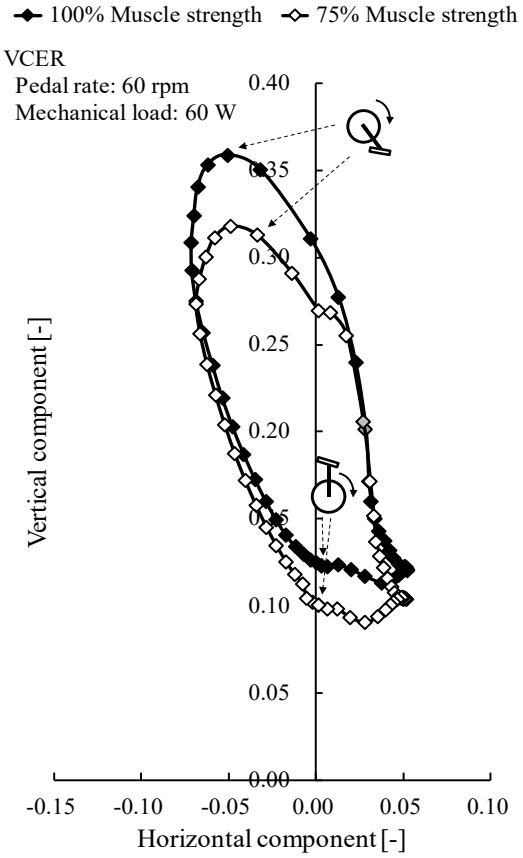


Figure 22. Comparison of pedal reaction forces between 100% muscle strength (filled symbols) and 75% muscle strength (unfilled symbols) of volitional conventional ergometer (VCER) exercise at 1 g, a pedal rate of 60 rpm and mechanical load of 60 W. Joint reaction force is shown in horizontal and vertical components, normalized by body weight. Gray symbol, lowest pedal position.

The resultant pedal reaction force at 0 g was lower than that at 1 g (Figure 11). The effective force during VCER exercise with a pedal rate of 60 rpm and a mechanical load of 60 W was less at 0 g than at 1 g (Figure 12). The effective and resultant forces at 0 g will be decreased by compensatory actions of the loss of function due to zero gravity to maintain a specific pedal rate against an external mechanical load. Activated muscles in the leg at 0 g (Figure 13) are associated with such compensation. Effective and resultant forces decreased when elderly persons pedaled on the ground at 1 g compared with younger persons under the same cycling conditions [24]. Knee joint reaction

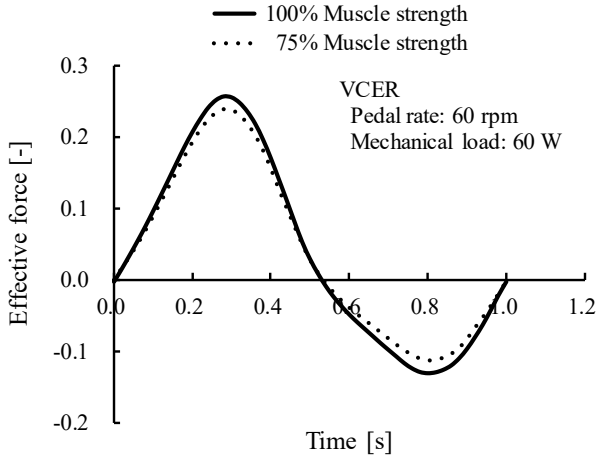


Figure 23. Comparison of effective force between 100% muscle strength (solid line) and 75% muscle strength (dotted line) of volitional conventional ergometer (VCER) exercise at 1 g, a pedal rate of 60 rpm and mechanical load of 60 W. Joint reaction force normalized by body weight is shown in horizontal and vertical components.

forces in elderly individuals (n = 9; age, 62–76 years) during ergometer cycling 15 months after TKR were measured [2]. Simulated VCER exercise at 1 g with leg muscles weakened to mimic elderly persons after TKR agreed with their experimental findings (Figure 5). Moreover, we confirmed the same decreasing trend in the kinetics of pedal reaction force in VCER exercise at 1 g with weakened muscles of the leg (Figures 22 and 23). Comparisons between experimental and simulation results can indicate different compensatory actions between VCER with full muscle strength at 0 g and reduced muscle strength at 1 g.

From the perspective of physical adaptation, it is interesting that decreases in the magnitude of effective and resultant forces showed an identical trend during ergometer cycling by elderly persons at 1 g and healthy persons at 0 g. A loss of function in external factors such as a weightless environment should be compensated to realize the same motion on the Earth. The internal factor of weakened muscles will require elderly persons to choose different compensation methods, such as muscle coactivation [25].

A compressive load along the bone axis coupled with some high-force impulsive loading is generally required to maintain bone health. The following effects of repetitions per day for six weeks using an identical load on the ulna of mature turkeys were indicated [26]: a minimal number of repetitions is required to prevent bone loss, a greater number of repetitions is required to gradually increase bone formation, and even with an increase in the number of repetitions, there is a specific number of repetitions that shows no significant difference in bone formation. From a response relationship between peak strain magnitude and changes in ulnar bone mass in mature turkeys, the minimal effective strain to prevent bone loss was identified [27]. The effects of disuse, adaptation, mild overload and pathological overload on the strength of load-bearing bones were associated with the magnitude of strain [28]. The effects of both the number of repetitions and of the magnitude of joint reaction forces caused by muscle contraction force, external force, and virtual gravity on body elements must be considered under microgravity and hypogravity.

Astronauts are at high risk for bone fracture, especially during exposure to the stresses of re-entry into the atmosphere of the Earth after a long period of weightlessness [3]. Ergometer exercise can reduce the load imposed by gravitational force on the leg joints with a metabolic cost comparable to that of walking on the Earth [2]. This type of exercise is beneficial for persons with knee osteoarthritis and elderly individuals planning to engage in cardiovascular training but not for persons who are exposed to long periods of weightlessness. Fortunately, the HTS is compact, and should become a useful countermeasure to induce coactivated muscle activity during ergometer exercise. It allows repeated high loading on bones during long-term weightlessness (Figures 15, 17 and 20), which will elicit adaptive bone remodeling.

The great trochanter becomes highly deconditioned during long space flights [29]. Young males have undergone unilateral lower leg suspension (ULLS) for four weeks in a model of a weightless environment [30,31]. We conducted a ULLS experiment that included a healthy man with one foot completely weight-free, and a ULLS+HTS experiment involving four healthy men. The HTS electrical stimulation was applied to the quadriceps and hamstrings for 16 min three times a week while seated. Deconditioning of the hip at the femoral neck and greater trochanter was prevented. We confirmed the activities of the abductor, gluteus medius and minimus muscles attached to the greater trochanter using electromyography. Muscle activities of the gluteus medius and minimus muscles were simulated at 0 g and 1 g during VCER exercise (Figure 13), which elicited more activity of the vastus lateralis and rectus femoris of the knee extensor at 0 g than at 1 g, whereas the activity of the hip abductor at 0 g considerably decreased. Applying HTS stimulation to the quadriceps and the hamstring at intensities of 10% and 20%, respectively, recovered mean muscle forces induced by the gluteus minimus and medius 4.5- and 2.5-fold, respectively, compared with the absence of HTS stimulation at 0 g. Much larger loading will improve the reduced mineral density of the greater trochanter at 0 g.

Centrifugal acceleration is a physical phenomenon that might serve as an alternative to gravitational acceleration, and it is referred to as artificial gravity. Because the primary factor affecting physical deconditioning during spaceflight is the loss of gravitational loading and stimulation, the most effective physical countermeasure would be to introduce gravity [3]. Thus, artificial gravity generated by centrifugal acceleration has been studied. However, joint reaction forces in a rotating human-ergometer have not been analyzed from the viewpoint of preventing bone loss and reduced muscle strength. We examined oxygen uptake and joint reaction forces with altered rotational motion of the RER model while cycling at a pedal rate of 60 rpm and a mechanical load of 60 W (Figure 18). The resultant force on each vertebral body (Figure 19) at a rotating velocity > 360 deg/s was close to the results of walking on a level surface at 1 g [21], independently of the effects of HTS stimulation of the leg muscle. The eccentric force of the soleus at higher artificial gravity far exceeded the lower boundary force of the soleus induced by the HTS. The HTS affected the level of reaction forces at all leg joints whereas artificial gravity did not affect those at the knee and hip joints (Figure 20). Stimulation of the ankle joint increased oxygen uptake during HCER exercise at 1 g (Figure 7) as well as reaction force on the ankle joint (Figure 20). Intensive contractions of the gastrocnemius and the soleus muscles induced by the HTS will prevent the loss of bone density in the calcaneus.

An optimal method enabled the determination from conflicting characteristics, of the pedal rate and mechanical load during ergometer cycling that generates exercise intensity and loads on leg joints equivalent to those extant during daily physical activities on the Earth. Optimal cycling conditions were simulated to generate specific values for oxygen uptake and knee joint reaction force at 0 g under the constraints of using specific values for oxygen uptake as an index of aerobic and resistive exercise intensity, and of the knee-joint reaction force required to maintain bone mass and strength. The validity of the cycling exercise model with the HTS for oxygen uptake at 1 g was shown by comparisons between simulated and experimental results (Figure 7). The simulated optimal cycling conditions showed that muscle co-contraction during the HTS cycling under microgravity, which decreases both the pedal rate and mechanical load on the ergometer pedal (Table 1), can offer a strategy to decide the most effective training conditions derived from various combinations. This is an advantage for the operational demands of space vehicles

or isolated bases located far from the Earth to maintain human health with time constraints and minimal environmental disruption.

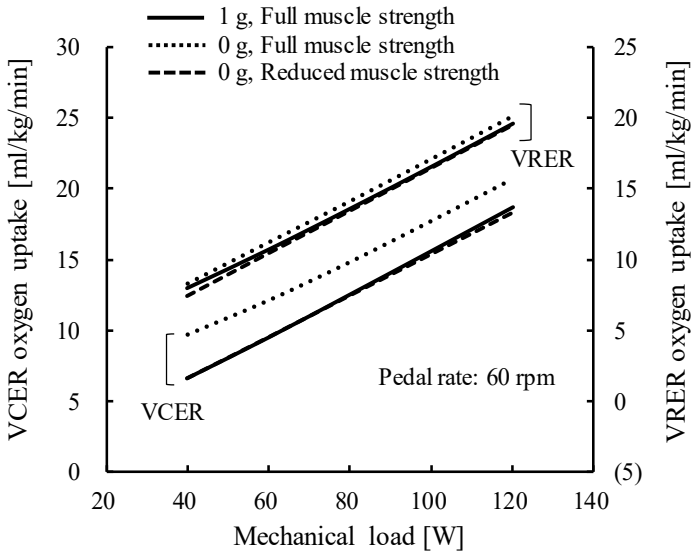


Figure 24. Comparison of oxygen uptake at 1 g and 0 g between volitional conventional ergometer (VCER) and volitional recumbent ergometer (VRER) exercise at full and reduced muscle strengths under different mechanical loads. Representative deconditioning in simulation with reduced muscle strength was assumed for muscles opposing gravity effects, and the muscle recruitment criterion was changed for the decrease in oxygen uptake. Pedal rate, 60 rpm.

Less oxygen is uptaken under experimental microgravity than on the Earth [20], which reflects lower metabolic internal power. The experiment was repeated during a 180-day flight aboard the Mir station. Long-term exposure to weightlessness resulted in physiological deconditioning, and crews underwent repeated measurements under deconditioning. The oxygen uptake in Figure 9 was simulated under a healthy physical condition. To compare with the experiment under long-term exposure to microgravity, weakened muscle was assumed in the VCER and VRER models as follows.

An oxygen uptake of 40 ml/kg/min in the simulation is approximately 84% of the mean maximal oxygen uptake in healthy adult males (age, 20-31 years) during ergometer exercise at 1 g [33]. Having 3-fold body weight in the simulated and desired reaction forces at the knee joint and the simulated reaction forces at the hip joint approximates the measured maximum magnitude of resultant forces at the knee joint during level walking in vivo at 1 g [34,35] and at the hip joint [35,36], respectively.

6. Conclusions

The validity of the ergometer exercise models was proven by comparisons with experimental findings. Model analysis provided perspectives even with widely varying parameters and altered exercise conditions. The HTS principle with exercise predicted an effective countermeasure to prevent deconditioning during lengthy exposure to microgravity. This approach will support future travel to Mars and will offer a benefit against hypogravity on the surfaces of the Moon and Mars. An optimal method could provide a better alternative to prevent deconditioning considering exercise intensity and loads on joints induced by daily activities at 1 g and operational demands in isolated equipment and facilities located far from the Earth.

This simulation was limited by the fixed leg angles and oxygen uptake, even under altered conditions of cycling and gravity. Leg angles during closed kinematic exercise such as cycling on an ergometer change according to age [24] and pedal rate [37]. The entire body including the cardiovascular and pulmonary systems undergoes deconditioning in microgravity [3], which alters the muscle fiber component ratio [38]. Open kinematic exercise such as knee and ankle joint flexion and extension does not require an ergometer or coordination between both legs during cycling. This exercise combined with the HTS will be simulated to confirm the amount of oxygen uptake and joint reaction forces. Changes in the ratio of the number of fast- to slow-twitch muscle fibers that are associated with oxygen uptake ability will be confirmed using the AMS, because these ratios increase under microgravity [38].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP26506014, JP265006025.

References

[1] Y. Tagawa, N. Yamamoto, M. Omoto, H. Matsuse, and N. Shiba, “Simulation of Oxygen Uptake and Leg Joint Reaction Force During Ergometer Exercise Under Altered Gravity,” in 39th Annual International Conference of IEEE EMBC, Jeju Island, Korea, 2017. DOI: 10.1109/EMBC.2017.8037219.

[2] I. Kutzner, B. Heinlein, F. Graichen, A. Rohlmann, A.M.Halder, A. Beier, and G. Gergmann, “Loading of the knee joint during ergometer cycling: telemetric in vivo data,” J. Orthop. Sports Phys. Ther., **42**(12), 1032-1038, 2012. DOI: 10.2519/jospt.2012.4001

[3] G. Clement and A. Bukley (Editors), Artificial Gravity, Microcosm Press and Springer, 2007.

[4] T. Yanagi, N. Shiba, T. Maeda, K. Iwasa, Y. Umezu, Y. Tagawa, S. Matsuo, T. Yamamoto, K. Nagata, and J.R. Basford, “Agonist contractions against electrically stimulated antagonists,” Arch. Phys. Med. Rehabil., **84**(6), 843-848, 2003. DOI: 10.1016/S0003-9993(02)04948-1

[5] N. Shiba, H. Matsuse, Y. Takano, K. Yoshimitsu, M. Omoto, R. Hashida, Y. Tagawa, T. Inada, S. Yamada, and H. Ohshima, “Electrically stimulated antagonist muscle contraction increased muscle mass and bone mineral density of one astronaut - Initial verification on the International Space Station,” PLoS One, **10**(8), 14 pages, 2015. DOI: 10.1371/journal.pone.0134736

[6] T. Weber, A.A. Al-Munajjed, G.J. Verkerke, S. Dendorfer, and T. Renkawitz, “Influence of minimally invasive total hip replacement on hip reaction forces and their orientations,” J. Orthop. Res., **32**(12), 1680-1687, 2014. DOI: 10.1002/jor.22710

[7] Under Interface Features, Lesson 2-4, The AnyBody Modeling System version 6.0.6, AnyBody Tutorials, 2016.

[8] E.T. Howley and B.D. Franks, Health Fitness Instructor's Handbook, Human Kinetics, 1997.

[9] J. Rasmussen, M. Damsgaard, and M. Voight, "Muscle recruitment by the min/max criterion –a comparative numerical study," J. Biomech., 34(3), 409-415, 2001. DOI: 10.1016/S0021-9290(00)00191-3

[10] M. Damsgaard, J. Rasmussen, S. T. Christensen, E. Surma, and M. de Zee, "Analysis of musculosketetal systems in the AnyBody Modeling System," Simul. Model Pract. Th., **14**(8), 1100-1111, 2006. DOI: 10.1016/j.simpat.2006.09.001

[11] M.L. Ohlsson and M. Gulliksson, "Least squares approach to inverse problems in musculoskeletal biomechanics," Technical report, Mid Sweden University, 10 pages, 2009.

[12] M. Omoto, H. Matsuse, Y. Takano, S. Yamada, H. Ohshima, Y. Tagawa, and N. Shiba, "Oxygen uptake during aerobic cycling exercise simultaneously combined with neuromuscular electrical stimulation of antagonist," J. Nov. Physiother., **3**(6), 185-191, 2013. DOI: 10.4172/2165-7025.1000185

[13] M. Tokui, "Analysis of the muscular mechanical efficiency focused on internal and external power," Ph.D Thesis, Kyushu Institute of Technology, 2008, (in Japanese).

[14] J.J. Seabury, A.C. Adams, and M.R. Ramey, "Influence of pedaling rate and power output on energy expenditure during bicycle ergometry," Ergonomics, **20**(5), 491-498, 1977. DOI: 10.1080/00140137708931658

[15] H.I. Griffiths, "Comparison of energy expenditure among three different cycling positions," Graduate Student These, Dissertations, & Professional Papers, University of Montana, 1989.

[16] Y. Tagawa and N. Yamamoto, "For improvement of motor function by surface electrical stimulation," Special Issue, J. Inst. Elect. Engnr. Jpn., **136** (19), 666-669, 2016, (in Japanese).

[17] T. Watanabe, Y. Tagawa, E. Nagasue, and N. Shiba, "Surface electrical stimulation to realize task oriented hand motion," in 31st Annual International Conference of the IEEE EMBS, Minneapolis, Minnesota, USA, 2009. DOI: 10.1109/IEMBS.2009.5333812

[18] T. Watanabe, Y. Tagawa, and N. Shiba, "Mapping method using a super multi-electrical stimulation device," in 15th IFESS Annual Conference, Vienna, Austria, 2010.

[19] M. Gobbo, N.A. Maffiuletti, C. Orizio, and M.A. Minetto, "Muscle motor point identification is essential for optimizing neuromuscular electrical stimulation use," J. NeuroEng. Rehabili., **11**(6), 6 pages, 2014. DOI: 10.1186/1743-0003-11-17

[20] M. Girardis, D. Linnarsson, C. Moia, D.R. Pendergast, and G. Ferretti, "Oxygen cost of dynamic leg exercise on a cycle ergometer: effects of gravity acceleration," Acta Physiol. Scand., **166**(3), 239-246, 1999. DOI: 10.1046/j.1365-201x.1999.00564.x

[21] A. Rohlmann, M. Dreischarf, T. Zandr, F. Graichen, and G. Bergmann, "Loads on a vertebral body replacement during locomotion measured in vivo," Gait & Posture, **39**(2), 750-755, 2014. DOI: 10.1016/j.gaitpost.2013.10.010

[22] S. Dendorfer and S. Torholm, "Harness – walking in microgravity," Final Report on Feasibility Study, Report no: 21385/08/NL/PA, AnyBody Technology A/S, 20-24, 2008.

[23] G.A. Dudley, P.A. Tesch, B.J. Miller, and P. Buchanan, "Importance of eccentric actions in performance adaptations to resistance training," Aviat. Space Environ. Med., **62**(6), 543-550, 1991.

[24] J.W. Seo, D.H. Kim, S.T. Yang, D.W. Kang, J.S. Choh, and G.R. Tack, "Comparison of joint kinematics and pedaling force in the young and the elderly," J. Phys. Ther. Sci., **28**(8), 2245-2248, 2016. DOI: 10.1589/jpts.28.2245

[25] Y. Tagawa, Y. Ogata, and N. Yamamoto, "Simulation of muscle activity during cycling with full and reduced muscle strengths of the leg," in 39th Annual International Conference of IEEE EMBC, Poster, Jeju Island, Korea, 2017.

[26] C.T. Rubin and L.E. Lanyon, "Regulation of bone formation by applied dynamic loads," J. Bone Joint Surg., **66-A**(3), 397-402, 1984.

[27] C.T. Rubin and L.E. Lanyon, "Regulation of bone mass by mechanical strain magnitude," Calcif. Tissue Int., **37**(4), 411-417, 1985.

[28] H.M. Frost, "A 2003 update of bone physiology and wolff's law for clinicians," Angle Orthod., **74**(1), pp.3-15, 2004. DOI: 10.1043/0003-3219(2004)074<0003:AUOBPA>2.0.CO;2

[29] M.R. Barratt and S.L. Pool (Editors), Principle of Clinical Medicine for Space Flight, Springer, 2008.

[30] T. Ito, Y. Tagawa, S. Tanaka, N. Shiba, Y. Umezu, T. Yamamoto, and J.R. Basford, "Development of practical and effective hybrid exercise for use in weightless environment," in 26th Annual International Conference of the IEEE EMBS, San Francisco, USA, 2004. DOI: 10.1109/IEMBS.2004.1404185

[31] R. Takeuchi, Y. Tagawa, T. Inada, K. Suetsugu, K. Mimura, and N. Shiba, "A countermeasure for preventing atrophy of musculoskeletal system under microgravity," in SICE-ICCAS 2006 SICE-ICASE International Joint Conference, Busan, Korea, 2006. DOI: 10.1109/SICE.2006.315451

[32] M. Praagman, E.K.J. Chawick, F.C.T. van der Helm, and H.E.J. Veeger, "The relationship between two different mechanical cost functions and muscle oxygen consumption," J. Biomech., **39**(4), 758-765, 2006. DOI: 10.1016/j.jbiomech.2004.11.034

[33] T. Matsuo, K. Ohkawara, S. Seino, N. Shimojo, S. Yamada, H. Ohshima, K. Tanaka, and C. Mukai, "An exercise protocol designed to control energy expenditure for long-term space missions," Aviat. Space Environ. Med., **83**(8), 783-789, 2012. DOI: 10.3357/ASEM.3298.2012

[34] I. Kutzner, B. Heinlein, F. Graichen, A. Bender, A. Rohlmann, A. Halder, A. Beier, and G. Bergmann, "Loading of the knee joint during activities of daily measured in vivo in five subjects," J. Biomech., 43(11), 2164-2173, 2010. DOI: 10.1016/j.jbiomech.2010.03.046

[35] P. Damm, I. Kutzner, G. Bergmann, A. Rohlmann, and H. Schmidt, "Comparison of in vivo measured loads in knee, hip and spinal implants during level walking," J. Biomech., **51**, 128-132, 2017. DOI: 10.1016/j.jbiomech.2016.11.060

[36] G. Bergmann, A. Bender, J. Dymke, G. Duda, and P. Damm, "Standardized loads acting in hip implants," PloS One, **11**(5), 23 pages, 2016. DOI: 10.1371/journal.pone.0155612

[37] R.R. Bini, M. Rossato, F. Diefenthaeler, F.P. Carpes, D.C. dos Reis, and A.R.P. Moro, "Pedaling cadence effects on joint mechanical work during cycling," Isokinet. Exerc. Sci., **18**(1), 7-13, 2010. DOI: 10.3233/IES-2010-0361

[38] V.R. Edgerton, M.Y. Zhou, Y. Ohira, H. Klitgaard, B. Jiang, G. Bell, B. Harris, B. Saltin, P.D. Gollnick, R.R. Roy, M.K. Day, and M. Greenisen, "Human fiber size and enzymatic properties after 5 and 11 days of spaceflight," J. Appl. Physiol., **78**(5), 1733-1739, 1995. DOI: 10.1152/jappl.1995.78.5.1733

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 4, No. 1, 08-16 (2019)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Spectral Grid Impedance Identification on the Low-, Medium- and High-Voltage Level – System Design, Realization and Measurement Results of Grid Impedance Measurement Devices

Hauke Wilken*, Michael Jordan, Detlef Schulz

Helmut-Schmidt-University, Chair of Electrical Power Systems, Holstenhofweg 85, 22043 Hamburg, Germany

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 30 December, 2018
Online: 12 January, 2019

Keywords:
Grid impedance
Power quality
Grid integration

ABSTRACT

The frequency dependent grid impedance at the terminals of an electrical supply grid is an essential parameter for power quality and grid feedback analysis. In this contribution the identification of the grid impedance is achieved using spectral excitation currents at the corresponding grid connection point, which generate measureable changes in the grid voltage spectra. Different measurement systems based on this method have been successfully realized for the low- and medium-voltage level. A measurement system for the high-voltage level is currently being realized. The application of the measurement systems are facing quite different external framework conditions leading to specific system designs for each measurement device. The spectral grid impedance identification on these voltage levels is done through fast switching of an ohmic load with power electronic. The requirements and setup of the grid impedance measurement systems are outlined and measurement results on the low- and medium-voltage level are presented and discussed. The systems can be used to evaluate grid connection points and to verify other active or passive grid impedance identification methods on different voltage levels.

1. Introduction

This paper is an extension of work originally presented in 2017 at the IEEE Power & Energy Society General Meeting [1]. The extensions consider essential aspects on the relationships of the grid impedance and the influence of harmonics on the grid voltages, especially in relation with renewable energy systems. Additionally the measurement results and possible benefits are described in more detail. The harmonic requirements are discussed in relation to the experimental results. Further on a new possible measurement procedure for an improved grid integration of renewable energy systems is presented. Finally the connection aspects and structure of the devices are explained in more detail.

The share of decentralized renewable energy systems will continue to grow in the upcoming years. Their share in the total electricity generation in Germany is targeted to rise up to 80 % by 2050. This changes the power flows in the grid, which have an effect on the grid voltage levels, the loads of the grid equipment and the protection technology. Therefore, a strategic grid reinforcement, grid optimization and grid expansion is required to master the future challenges.

The electrical characteristics of a point of common coupling (PCC) are determined by the grid voltage and the internal impedance at the PCC. At the fundamental frequency these parameters define the short circuit power and hence the capacity of the PCC. To evaluate harmonics above the fundamental frequency of power systems connected to the PCC, the spectral characteristic of both parameters are needed. In most cases the grid voltages can be easily measured and spectrally analyzed. The grid impedance at the fundamental frequency can be approximately obtained from grid simulation programs. It is typically not possible to determine the grid impedance above the fundamental frequency through simulation because this requires detailed spectral models of the power system utilities. The spectral models however are often not available or just valid for low frequencies. Therefore an exact determination of the spectral grid impedance over a wide spectral range can only be accomplished by measurement.

The frequency-dependent grid impedance can be used to evaluate network perturbations (harmonics, flicker) [2]. At the Helmut Schmidt University, measurement devices for the determination of the time- and frequency-dependent grid impedances at the low-, medium- and high-voltage level have been developed [3], [4], [5].

2. Relevance and interrelation of the grid impedance

The grid impedance is a complex variable that describes the resulting voltage uplift or voltage drop when power is fed into the

*Hauke Wilken, +49 40 6541-2329, hauke.wilken@hsu-hh.de

grid or power is taken from the grid. It is composed of the proportions of the individual impedances of the used grid equipment. In the following, the 50 Hz component of the grid impedance of a grid connection point / point of common coupling (PCC) in the medium-voltage grid is described. It essentially consists of the higher-level 110 kV grid, the 110/10 kV transformer and the medium voltage cable, shown in Fig. 1. For this purpose, one ohmic and one inductive component are used.

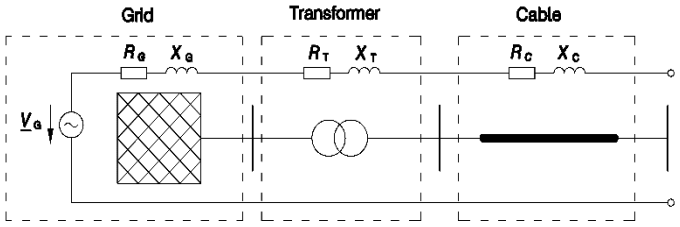


Figure 1. Schematic of grid equipment

2.1. Superordinate 110-kV-Grid

The proportion from the higher-level 110 kV grid can be described with the terminal voltage $\underline{V}_{nG}/\sqrt{3}$ as:

$$\underline{Z}_G = R_G + jX_G = \frac{1{,}1 \cdot V_{nG}}{\sqrt{3} \cdot \underline{I}_k''} \tag{1}$$

According to DIN EN 60909, a safety margin of 10 % is taken into account for the short-circuit current $\underline{I}_k''$. The impedance $\underline{Z}_G$ is related to the short circuit apparent power S_k''. The amount can be stated, with usual values according to Table 1, to:

$$|\underline{Z}_G| = \frac{1{,}1 \cdot V_{nG}^2}{S_k''} \tag{2}$$

Table 1: Typical short-circuit currents and short-circuit powers [6]

V_{nG}	I_k''	S_k''
10 kV	29 kA	0,5 GVA
110 kV	42 kA	8 GVA
220 kV	63 kA	24 GVA
380 kV	80 kA	53 GVA

The relevant reactance X_G can be equated approximately with the impedance Z_G for the high voltage level. Since the transformer in most cases has a step control, the regulated voltage value of the low-voltage side of is used for V_{nG}. The reactance / resistance ratio can be estimated with a factor of 6 if unknown [7].

2.2. 110/10-kV-Transformer

The impedance of the transformer can be calculated for the middle position of the step control with the relative short-circuit voltage of v_k and copper losses of P_{Co} to:

$$X_T \approx Z_T = \frac{V_{nG}^2}{S_{kT}} = \frac{V_{nG}^2}{\frac{S_{rT}}{v_k}} \tag{3}$$

$$R_T = \frac{V_{nG}^2 \cdot P_{Co}}{S_{rT}^2} \tag{4}$$

2.3. Medium-voltage cable

The impedance of the medium-voltage cable depends on the corresponding length as well as the resistance coating R_C' and reactance coating X_C'.

2.4. Grid impedance at the point of common-coupling

The impedance Z_{PCC} at the grid connection point / the point of common-coupling (PCC) results from the sum of the individual impedances. It is related to the short-circuit power $S_{k,PCC}$ and the voltage V_{PCC} as follows:

$$S_{k,PCC} = \frac{V_{PCC}^2}{Z_{PCC}} \tag{5}$$

Fig. 2 shows the grid impedance as well as the short circuit apparent power of a PCC as a function of increasing cable length. It shows the indirect proportionality between $S_{k,PCC}$ and Z_{PCC}. This relationship plays the decisive role in the calculation of the permissible voltage change by connection of loads or generation units.

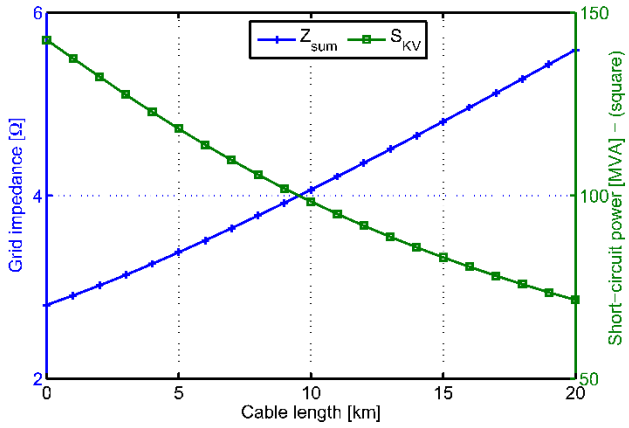


Figure 2. Characteristic of grid impedace (solid) and short-circuit power (dashed)

2.5. Voltage change

According to the technical guideline for the connection of power generation units in Germany, the voltage change caused at the PCC Δv_a in the medium-voltage grid must not exceed 2 % [7]. Using the power ratio of $S_{k,PCC}$ to the apparent power of the power generation unit, the voltage change can be roughly estimated. With a ratio smaller than 50, a closer look is necessary, since the power generation unit has more influence on the grid voltages. For this, the grid impedance angle as well as the injected reactive current must be taken into account, which can be set at the medium voltage level via the power factor between $\cos\varphi_a = 0{,}95_{ind}$ and $\cos\varphi_a = 0{,}95_{cap}$ (inductive / capacitive power feed-in and over- / under-excited operation respectively). The voltage change is a function of the grid impedance angle ψ_{PCC} at the PCC, the maximum system apparent power $S_{a,max}$, the grid impedance $\underline{Z}_{PCC}$ and the voltage V_{PCC}.

$$\Delta v_a = \frac{S_{a,max} \cdot \cos(\psi_{PCC} + |\varphi_a|)}{S_{k,PCC}} \tag{6}$$

Solving this relationship to the maximum apparent power of the generation unit that can be operated at the corresponding PCC, and using the ohmic and inductive components of the grid impedance Z_{PCC}, yields:

$$S_{\mathrm{a,max}} = \frac{\Delta v_{\mathrm{a}} \cdot V_{\mathrm{PCC}}^2}{|R_{\mathrm{PCC}} \cdot \cos(\varphi_{\mathrm{a}}) - X_{\mathrm{PCC}} \cdot \sin(\varphi_{\mathrm{a}})|} \quad (7)$$

It can be observed that for a power feed-in with a power factor of $\cos \varphi_{\mathrm{a}} = 1$, only the resistive component is relevant for the voltage uplift.

3. Grid feedback

There are a number of standards for ensuring the proper operation of electrical power generation units and electrical loads connected to the power grid. At the PCC, the grid voltages must meet certain criteria that are understood by the grid quality. Failures and outages of the energy supply grid (long and short-term outages) fall under the security of supply. In reference with the grid impedance, in particular the voltage form and the influence of generating units and consumers on the grid voltages are of interest. The grid voltages should be ideally sinusoidal and have a constant frequency (50 or 60 Hz). Grid perturbations due to harmonics (also intermediate and subharmonic), voltage fluctuations and unbalances as well as transients (also voltage dips and overvoltages) reduce the grid quality. Grid perturbations result from non-linear equipment and switching operations. Especially the increasing proportion of power electronics in many devices and also renewable energy systems can lower the power quality by taking or feeding in non-ideal sinusoidal currents. The permitted limits for the harmonic currents and also for the harmonic voltages must not be exceeded [8]. Harmonics can have several bad effects in the electrical energy supply network:

- Reduction of efficiency
- Destruction of insulation
- Reduction of lifetime (early failure of insulation)
- Thermal overload of equipment
- Malfunction of protection and measuring devices
- Attenuation and distortion of signals (eg frequency remote control, grid communication…)

As the number of power electronics within consumers and power generation units connected to the electrical energy supply grid increases, the effects increase. With measured frequency characteristics of the grid impedances, the grid perturbations can be better and more objectively evaluated. Resonances in the frequency characteristic of the grid impedances can be detected at an early stage.

In general, high short circuit power and hence a small grid impedance value results in lower voltage drops when connecting loads, and also in a low voltage uplift when connecting power generation units as well as in reduced grid perturbations.

Table 4 summarizes limits for the allowable harmonic voltages at the low, medium and high voltage levels. The harmonic limits of standard EN50160 describe the minimum requirements for PCCs. They are therefore higher than the values of the technical guidelines for power generation units, including renewable energy systems ([7], [9]). Deviating from these limitation values can be tolerated by grid operators with higher values in subareas of their electrical grids in coordination with customers if a negative influence on consumers and power generation plants can be avoided.

Table 2: Selected limitation values for certain harmonics at low-, medium- and high-voltage level

Order ν	Vν/Vn at low- and medium-voltage level (acc. to [8])	Vν/Vn at medium-voltage level (acc. to [7])	Vν/Vn at high-voltage level (acc. to [8])	Vν/Vn at high-voltage level (acc. to [9])
5	6.0 %	0.5 %	5.0 %	0.25 %
7	5.0 %	1 %	4.0 %	0.5 %
11	3.5 %	1 %	3.0 %	0.5 %
13	3.0 %	0.85 %	2.5 %	0.4 %
17	2.0 %	0.65 %	tba.	0.3 %
19	1.5 %	0.6 %	tba.	0.25 %
23	1.5 %	0.5 %	tba.	0.2 %
25	1.5 %	0.4 %	tba.	0.15 %

When connecting high power wind farms via HVDC systems to the high-voltage level (especially in the offshore area), the harmonic emissions of the entire system must be considered up to a frequency of 10 kHz [9]. In addition, the THD (Total Harmonic Distortion) is used to assess the power and voltage quality of renewable energy systems:

$$THD_{\mathrm{V}} = \frac{\sqrt{\sum_{i=2}^{40} V_{\mathrm{i}}^2}}{V_1} \quad (8)$$

At the medium voltage level, the THD_{V} must be below 8 % [7]. At the high voltage level, no limit value is specified for the THD_{V} in [9] and [8]. Furthermore, limitation values for the short- and long-term flicker must be complied, which were especially relevant for older directly coupled wind energy systems and have higher values than current systems using inverters.

The basis of the limitation values for harmonic voltages and currents is the assumption that the grid impedance consists of an ohmic and an inductive part, as described in section 2.1. However, capacitive elements in the grid equipment as well as non-linear behavior of consumers and power generation units with power electronics lead to a strong deviation from this assumption and to the occurrence of resonance frequencies. At these resonant points the limitation values can be exceeded due to power feed-in of renewable energy systems. Even instabilities can occur. In this case, the power must be reduced and possibly additional filters must be installed.

The design of grid-side and internal filters can be improved with knowledge of the grid impedance [10]. The internal control of the voltage and current regulators used in the inverters can also be adapted to the network impedance to avoid possible current fluctuations and instability problems [11].

4. Basic Principle of applied Grid Impedance Measurement Method

Several methods have been introduced to measure the grid impedance. They can be categorized into invasive and non-

invasive techniques. Non-invasive techniques assess natural power flow variations caused by fluctuating loads or generators with statistical methods. An example of a grid impedance estimation with a statistical methods can be found in [12]. However, in most cases a high precision over a wide frequency range cannot be achieve with these methods.

Commonly used invasive techniques inject harmonic or interharmonic current signals to a PCC. In [13] an inverter system is used to measure the grid impedance on the low-voltage level by sweeping the frequency of the injected current signals in the range from 100 Hz to 10 kHz. Inverter based grid impedance measurement methods are applied almost exclusively on the low-voltage level. To measure the grid impedance at medium- or high-voltage level with this method, the inverter systems need to have very high output power. Further on, the inverter system is coupled with a special transformer to the grid. Transformers with high rated power that offer the necessary broad frequency range are not available on the market and have to be specially designed, which makes this method complex and cost-intensive. On the medium- and high-voltage level active switching of transformers [14] or capacitor banks [15] have been commonly used to identify the grid impedance with good results. The switching has been done by circuit breakers with a number of maximum cycles of operations, which strongly reduces the possible repetition rate. Further on, the duration of the generated excitation pulse signals are very short and the pulse amplitude depends on the switching moment, which cannot be controlled accurately with the circuit breakers. Therefore the spectral excitation of these current pulses is only sufficient at certain frequencies and cannot be varied in practice. These drawbacks can be overcome by generating pulsed current signals with power electronic switches and ohmic loads.

As shown in Fig. 3 a power electronic switch is used to pulse an ohmic load in order to excite the grid. The advantage of this method compared to switching loads with circuit breakers lies in the flexibility of the pulse patterns that can be applied to the power electronic switch. The voltages and currents of the first interval where the switch is open (subscript 1) and the second interval where the load is pulsed by the switch (subscript 2) are transformed into the frequency domain with FFT-algorithms. In the frequency domain the complex spectra of the grid impedance $\underline{Z}_G(\omega)$ can be calculated as follows:

$$\underline{Z}_G(\omega) = \frac{\underline{V}_1(\omega) - \underline{V}_2(\omega)}{\underline{I}_1(\omega) - \underline{I}_2(\omega)} = \frac{\Delta\underline{V}(\omega)}{\Delta\underline{I}(\omega)} \qquad (9)$$

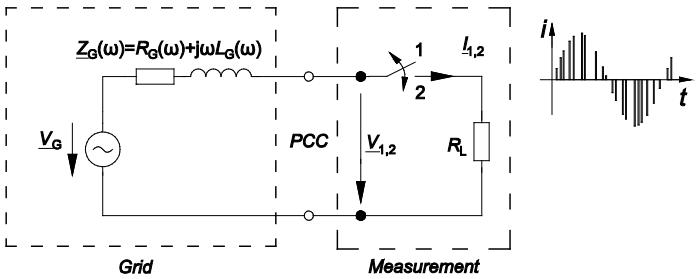


Figure 3. Grid impedance measurement with pulsed ohmic load

Strategies to excite the grid with pulsed current sequences can be found in [16], [17] and [18]. With periodic pulse signals the grid impedance can be precisely measured at specific frequencies. Spectral ranges of the grid impedance may be measured with a frequency sweep of periodic pulse signals. Depending on the spectral range and number of steps this method can be very time consuming. Another option is to switch the load with random pulse patterns. The resulting current signals have broad frequency spectra but with comparatively low amplitudes. In order to obtain good measurement results several measurement iterations in succession using random signals can be used.

On the low-voltage level with a four-wire system the pulsed current signals can be generated sequentially between two conductor lines or between a conductor line and the neutral line to measure the corresponding loop impedance parameters. Further on, equivalent line impedance parameters can be calculated out of the loop impedance parameters [18]. In three-wire systems as on the medium- and high-voltage level the loop and equivalent line impedance parameters can be measured in the same way. It is also possible to determine coupling impedance parameters between the lines with this method [4].

5. System Design and External Technical Parameters for the Realization of the Measurement Devices

The most important relevant technical parameters for the realization of the grid impedance measurement devices are given by the structure of the grids:

- Nominal voltage
- Grid structure
- Typical short-circuit powers

These parameters have to be compromised with the application and the definition of the subsequent measurement results:

- Frequency range of interest (only 50 Hz or spectral range)
- Repetition rate of measurements (in order to evaluate time changes)
- Measurement accuracy
- Framework conditions:
    - Price
    - Transportation
    - Grid connection

A. Technical Requirements

The power electronic switch is directly connected to the terminals of the different voltage systems. At the low-voltage level (0.4 kV) the effort is relatively low since there are single power electronic components e.g. Mosfets or IGBTs that fully offer the necessary blocking voltage. On the medium-voltage level (20 kV) a series connection of multiple power electronic components is required in order to achieve the high blocking voltages. At the high-voltage level (110 kV) there is an enormous effort. An extremely high demand of power electronic devices is necessary.

Another aspect concerns the spacing between parts within the measurement devices which are subjected to the grid voltages and other parts that are connected to ground, refer to Table III. At the low-voltage level a compact design on a printed circuit board is possible. At the medium- and high-voltage level this is not possible

due to the high grid voltages. The breakdown voltage in air amounts to about 30 kV/cm. A homogenous design however cannot be realized, so that a typical design value of 1 kV/cm can be used for an air-insulated design. The measurement system for the medium-voltage level has been realized with an air-insulated setup. At the high-voltage level distances of over 1 m would have to be held, making a compact air-insulated design impossible. Therefore ester is used as insulation medium.

Table 3: Overview of selected permitted isolation distances in air according to EN 61936-1

Highest voltage of facility	Minimum distances (line to earth and line to line)
12 kV	90 mm (indoor) 120 mm (indoor) 160 mm (indoor)
24 kV	160 mm 220 mm 270 mm
123 kV	900 mm 1100 mm

B. Grid Connection

The grid connections of the measurement systems are realized by circuit breakers so that possible short-circuits within the devices can be disconnected. On the one hand the dimensions and costs of the measurement devices for the different voltage levels strongly deviate. On the other hand the effort and costs for the connection to the grid also deviate severely. At the low-voltage level the power switch is very compact and grid connection can be done with simple cables and connectors. At medium- and high-voltage level the design, weight and laying of the cables as well as the necessary mountings are more complicated. Mounting of the sealing ends has to be done by specialized technical personnel. Especially at the high-voltage level the effort and costs are extremely high.

C. Organization Aspects

Dimensions and weight of the different grid impedance measurement systems vary to each other. At first the transportation of the systems differentiates a lot. The device for the low-voltage level can be easily carried. The measurement system for the medium-voltage level requires a small truck with a crane for transportation. The very large measurement container for the high-voltage level needs a heavy-weight transportation and for its placing a special crane. Additionally the area for assembly has to be considered. The measurement system for medium- and high-voltage level have to be put on solid ground, so that in some cases a fundament is needed. In order to connect these measurement devices to the grid the personnel requires qualification to obtain the authorization to connect and disconnect the devices to the grid.

6. Practical Realizations

Based on the method stated in section 4, measurement systems for the low- and medium-voltage level have been realized to measure the spectral grid impedance. Further on, a measurement system for the high-voltage level is currently being realized.

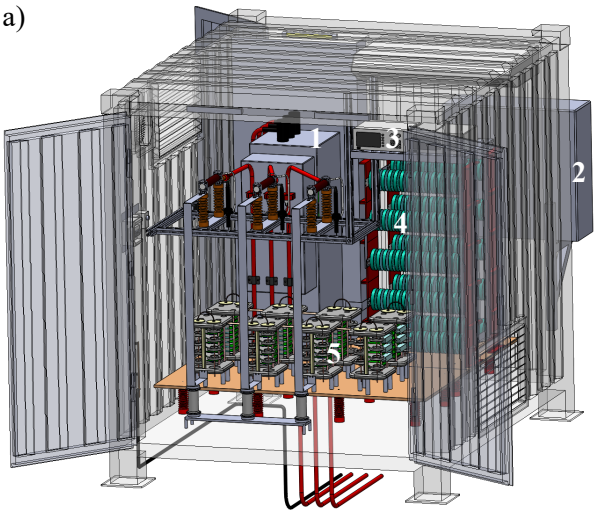
A. Low-Voltage Level (0.4 kV)

Fig. 4 shows the grid impedance measurement system for the low-voltage level. The system can measure the complex loop impedance parameters $\underline{Z}_{ab}(j\omega)$, $\underline{Z}_{bc}(j\omega)$ and $\underline{Z}_{ca}(j\omega)$ as well as the line impedance parameters $\underline{Z}_{a}(j\omega)$, $\underline{Z}_{b}(j\omega)$ and $\underline{Z}_{c}(j\omega)$ in four-wire systems with the outer conductor lines a, b and c up to 0.4 kV in the frequency range from DC to 150 kHz. The system generates asymmetrical pulsed current signals between two lines for the measurement, whereby every line combination can be pulsed sequentially. The maximum amplitude of the pulsed current signals can be varied between 1 A_{peak} to 20 A_{peak}.



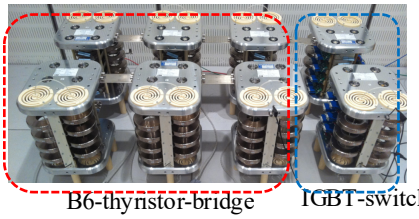
Figure 4. Grid impedance measurement system for the low-voltage level

a)



1) SF6- circuit-breaker (20 kV)
2) Air conditioning
3) Measurement and control units
4) High power resistors
5) Power electronic

b)



c)



Figure 5. 3D CAD model (a), photo of the power electronic grid excitation cirucit (b) and photo of the medium-voltage measurement container

B. Medium-Voltage Level (20 kV)

The setup of the grid impedance measurement system for the medium-voltage level is illustrated as a 3D-CAD model in Fig. 5 a). Fig. 5 c) shows a photo of the system. All components have been installed within a type-tested container. Additionally, to the power electronic circuit (Fig. 5 b)) and the load resistor a circuit-breaker, highly precise voltage and current transducers, high-voltage fuses, overvoltage suppressors, an air-conditioning system and a control and data recording system are installed in the system.

In most cases medium-voltage grids are operated with isolated or compensated starpoints. There is no neutral line/point explicitly available to identify the grid impedances. Therefore the method used on the low-voltage level is adapted, so that only the three conductor line impedance parameters are determined [19]. The system can measure the complex loop- and line impedance parameters in three-wire systems up to 20 kV_{rms} in the frequency range from DC to 20 kHz. The measurement device generates asymmetrical pulsed current signals between two outer conductor lines, whereby every combination can be sequentially pulsed. The maximum pulse amplitude can be varied from 1 A_{peak} up to 100 A_{peak}. The measured loop impedance parameters can then be rearranged to the equivalent impedance parameters of each outer conductor line:

$$\underline{Z}_a(j\omega) = \frac{1}{2} \cdot \left[\underline{Z}_{ab}(j\omega) - \underline{Z}_{bc}(j\omega) + \underline{Z}_{ca}(j\omega)\right] \quad (10)$$

$$\underline{Z}_b(j\omega) = \frac{1}{2} \cdot \left[\underline{Z}_{ab}(j\omega) + \underline{Z}_{bc}(j\omega) - \underline{Z}_{ca}(j\omega)\right] \quad (11)$$

$$\underline{Z}_c(j\omega) = \frac{1}{2} \cdot \left[-\underline{Z}_{ab}(j\omega) + \underline{Z}_{bc}(j\omega) + \underline{Z}_{ca}(j\omega)\right] \quad (12)$$

The power electronic grid excitation circuit is shown in Fig. 6. The B6-thyristor-bridge allows to select two outer conductor lines at a time. Then the IGBT-switch is pulsing the load resistor in order to excite the grid in the frequency region of interest. Each thyristor consists of 8 thyristors in series. The IGBT-switch is build up out of 16 IGBTs connected in series. The high power resistor can be varied in its resistance value in order to adapt the grid excitation current to the short-circuit power at the measurement point.

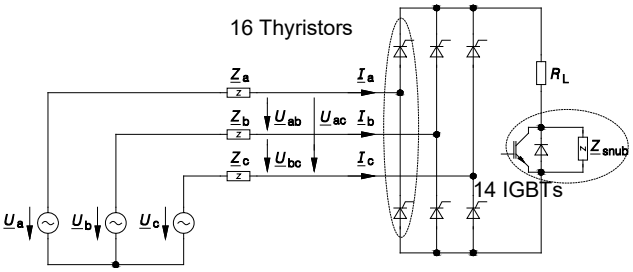


Figure 6. Grid impedance measurement circuit on the 20 kV medium-voltage level

C. High-Voltage Level (110 kV)

The grid excitation circuit for the high-voltage level is shown in Fig. 7. In comparison to the medium-voltage level there is a B6-diode bridge in combination with an IGBT and a resistive load. The B6-diode-bridge offers a simpler design and does not require control signals. Therefore the selection of the outer conductor lines has to be realized by the power switch.

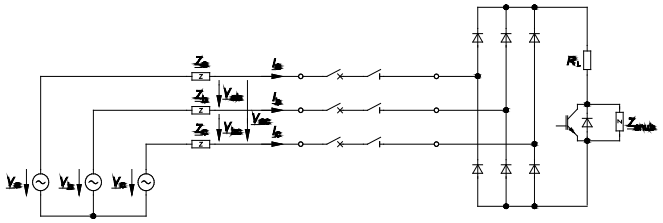
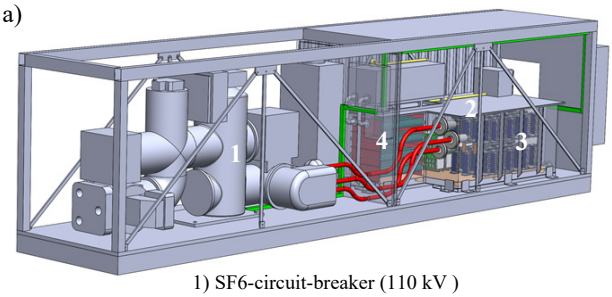


Figure 7. Schemetic circuit diagram for the gird impedance measurement on the 110 kV high voltage level

Fig. 6 shows a 3D-CAD model and photos of the current construction status. The power electronic circuit and the load are installed within a special tank filled with Ester as insulation medium. Each diode in the schematic diagram consists of 6 diode-stacks with each having 12 diodes in series. In total there are 432 diodes in the tank. The IGBT is built up of 15 stacks with each stack containing 6 IGBTs (90 in total). The peak current of the power electronic amounts to 200 A resulting in a possible peak grid excitation current of 22 MW. The temperature rising has to be observed and compensated by an air-conditioning system. The number and time length of the grid excitation correlates with the thermal energy and can be controlled by the measurement software system.

a)



1) SF6-circuit-breaker (110 kV)
2) Tank filled with ester as isolation medium
3) Power electronic modules
4) High power resistors

b)



Figure 8. 3D CAD model of 110 kV grid impedance measurement system (a) and photos of the current construction status (b)

Basic specifications of the measurement systems for the low-, medium- and high voltage level are stated in Tab. I. It can be seen that dimensions, weight and costs are highly deviating.

7. Measurement Results and Relevance for Grid Integration

On weekdays, there are generally noticeably different frequency profiles of the grid impedances than on weekends. There are also fluctuations between midday, evening and night hours. For the evaluation of grid perturbations, the frequency dependent grid impedance below 2 kHz (40th harmonic) is especially relevant. The

attenuation and frequency of the resonance points can vary greatly over time. At low attenuation, the parallel resonance has a lower frequency width but higher magnitude and at high attenuation vice versa. The damping is mainly determined by the local electrical loads. For Flicker, the 50 Hz grid impedance is crucial, where a single measurement is sufficient, since this usually varies only slightly in time. For the evaluation of harmonics a measurement of the frequency dependent grid impedance over several days or partial weeks is necessary. In particular, if temporally fluctuating feed-in capacities of renewable energies are to be investigated.

Table 3: Specifications of the grid impedance measurement systems for different voltage levels

Parameter	Low-Voltage Level (0.4 kV)	Medium-Voltage Level (20 kV)	High-Voltage Level (110 kV)
Dimensions (L×W×H)	0,5×0,4×0,2 m	2,9×2,5×2,5 m	12,2×2,5×2,7 m
Weight	7 kg	3,65 t	~26 t
Excitation current	1 – 20 A	1 – 100 A	200 A
Device costs	~8.000 €	~200.000 €	~2.500.000 €
Connection costs	0 €	~10.000 €	~500.000 €

A. *Low-Voltage Level (0.4 kV)*

Fig. 9 shows the frequency characteristics of the loop impedances between the conductor lines (“a”, “b” and “c”) and the neutral line (“N”) measured at a low-voltage PCC. The absolute values are increasing with frequency and show some weak parallel and series resonances. Due to a high number of single phase loads there are strong asymmetries in the characteristics. Up to a frequency of about 30 kHz the frequency characteristics are quite similar, but at higher frequencies especially the phase angles show large differences. Some resonances vary over time caused by load profiles, which especially change the damping of resonances. This effect is illustrated in Fig. 10.

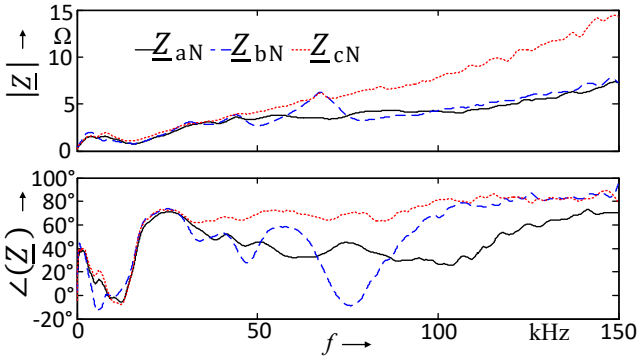


Figure 9. Absolute values and phase angles of the impedances between the conductor lines (“a”, “b” and “c”) and neutral line (“N”) measured at a low-voltage PCC

The first two parallel and series resonance points are below 3 kHz and are more attenuated during the day than at night. Reason could be a higher number of switching power supplies on the grid and thus higher cross-capacity due to computers and office equipment. Figure 10 shows the temporal variation of the first parallel resonance at approximately 1.6 kHz. It results from the interrelation of transformer and line inductances as well as capacities of cables and loads.

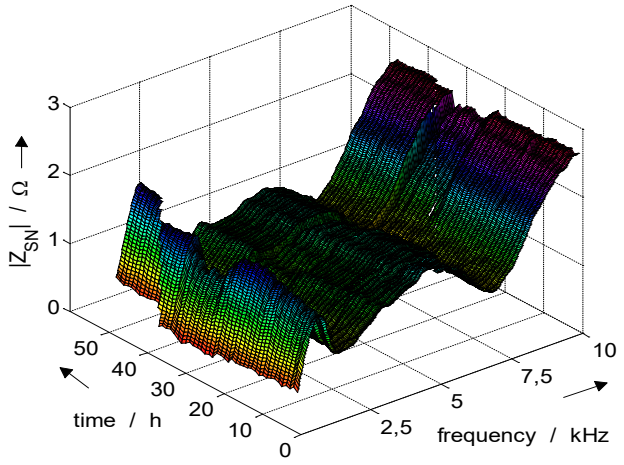


Figure 10. Absolute value of the loop impedance |ZaN| (conductor line “a” to neutral “N”) hourly measured over 3 days at a low-voltage PCC

B. *Medium-Voltage Level (20 kV)*

Fig. 11 presents the frequency characteristics of the loop impedance parameters measured at a medium-voltage PCC (20 kV). Only small asymmetries appear in the characteristics due to the geometric construction of transformers and power lines. The variation in time and frequency of the positive sequence component is displayed in Fig. 12. Two parallel resonance points can be seen at about 3.8 kHz and 12 kHz with high absolute values of about 400 Ω and 300 Ω, where the phase angle amounts to 0°. Series resonances can be seen at about 7.5 kHz and 15.5 kHz with local minimum impedance values and phase angles of 0°.

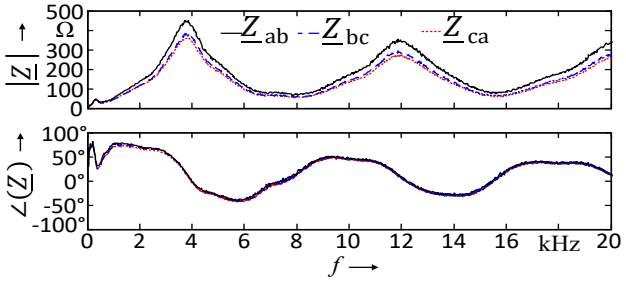


Figure 11. Absolute values and phase angles of the loop impedances between the conductor lines (“a”, ”b” and “c”) at a 20 kV medium-voltage PCC

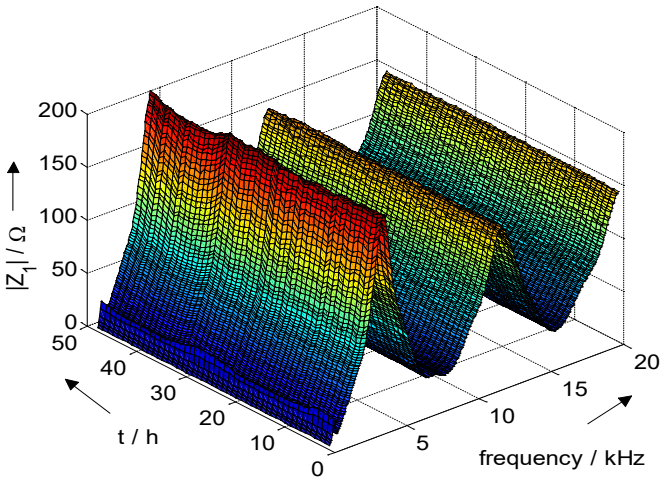


Figure 12. Absolute value of the positive sequence component |Z1| from DC to 20 kHz hourly measured over two days at a 20 kV medium-voltage PCC

Especially resonances in the frequency range below 1 kHz change over time with a day-night cycle. This effect can be seen in more detail in Fig. 13. The first parallel resonance point at 250 Hz (5th harmonic) shows higher peaks at night and more attenuation with less absolute values in the daytime.

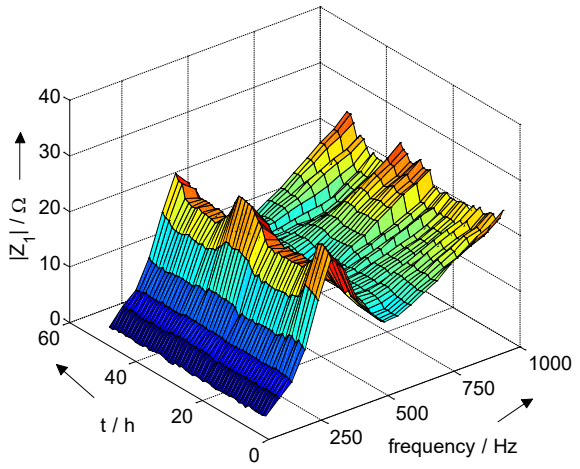


Figure 13. Absolute value of the positive sequence component |Z1| from DC to 1000 Hz hourly measured over two days at a 20 kV medium-voltage PCC

C. Relevance for Grid Integration

The measurement of the grid impedance can contribute to an improved grid integration of renewable energies. Its knowledge allows a more accurate calculation of the harmonic propagation in the grid. Currently, estimation methods based on measured RMS values of the injected currents are mostly used [2], [7], [9]. But when connecting renewable energy systems to the grid, two sources of harmonics should be considered in general. On the one hand, the renewable energy systems feed harmonic or inter-harmonic currents into the grid. On the other hand, the existing pre-load of the grid with harmonic voltages must be considered. With the frequency-dependent grid impedance and the complex values of the injected currents, the harmonic emission of renewable energies can be determined more accurately. Early detection of resonant frequencies and locations in the frequency characteristics of the grid impedances at the PCC can be used to better select grid equipment when connecting renewable energies. The knowledge of resonance points at the PCC is decisive for the determination of the pulse frequency in the used inverter. It can be varied within certain limits, whereby problems with excessive harmonic voltages at the affected frequencies can be avoided. Operation with variable pulse frequencies can reduce the dependence of the harmonic content on the operating point. Furthermore, the internal power and harmonic controllers can be better adapted to the grid. Power fluctuations and instabilities can be avoided. A lowering of the harmonic voltages can be achieved.

Additionally existing and possibly additionally required filter systems, as well as the renewable energy systems themselves and their inverters, can be designed more efficiently for the PCC. In this way, grid-side and internal filter systems and current controller loops can be better adapted to the PCC in advance as well as in the case of existing renewable energy systems. The dimensioning of a passive filter presupposes knowledge of the grid impedance or its assumed knowledge as well as the harmonic content of the grid. Temporal variations must also be considered in order to avoid resonances, as this can make the overall system unstable. The grid impedance is also needed for the design of active filters and their stable operation.

Furthermore, the maximum possible feed-in power of a PCC can be determined, as a result of which possibly additional systems can be connected to it. In summary an improved grid integration of renewable energy systems could look like this:

1. Measure the source voltages at the PCC with the identification of the existing harmonic voltage spectra
2. Measurement of the frequency dependent grid impedances with the developed grid impedance measurement systems
3. Identification of resonance points
4. Determination of the maximum grid connection capacity
5. Establishment of a mathematical model of the grid
6. Selection and dimensioning of suitable cables and equipment for connection of renewable energy systems
7. Simulation of the grid connection model and the plant certificate or if available a harmonic model of the renewable energy systems and the selected equipment
8. Detection of possible problems at resonance frequencies and, if necessary, adaptation of the pulse frequency of the inverter
9. Optimized design of filter systems for low harmonics
10. Optimized design of current controllers used within the inverter to avoid stability problems
11. Connection of additional renewable energy systems if limitation values have not yet been reached
12. Summary

This paper presents the challenges and the realization of three measurement devices for the grid impedance identification on the low-, medium- and high-voltage level. Their measurement results allow to determine the grid capacity of points of common coupling (PCC). The grid impedance at the fundamental frequency is directly related to the short-circuit power, so that conclusions about the tolerable connection power for loads or generation units can be stated. Then the measured frequency characteristic can be used to assess the grid feedback and propagation of harmonics especially within the context of an increasing share of renewable energies and their effect on the grid voltages. At first the used measurement method based on the pulsing of a known load using a special power electronic switch concept is described. The grid is excited in the frequency region of interest and the generated current and voltage transients are used to calculate the grid impedance. Then the applications of this method to the low-, medium- and high-voltage level are outlined. Due to the different voltages and short-circuit grid powers as well as grid structures there are quite different concepts necessary. Based on the experience of the small device for the low-voltage level, a 10-feet measurement container for the medium-voltage level has been built. Additionally the development of a 40-feet measurement container for the high-voltage level is introduced. These devices are compared with each other. Finally measurement results on the low- and medium-voltage level conclude this contribution. An improved grid integration of especially renewable energy systems can be enabled. With these novel devices a validation of other grid impedance measurement devices based on different methods can be achieved.

In the future the miniaturization of the presented measurement devices will be further investigated and pursued.

Acknowledgment

The projects “Development of a measurement device for the determination of the time and frequency dependent grid impedance on the medium-voltage level” and “Development of a measurement device for the determination of the frequency dependent grid impedance on the high voltage level up to 110 kV for the assessment of the availability of grid capacities as a system parameter for the dimensioning of energy storages” was/is funded by the German Federal Ministry for Economic Affairs and Energy under the support codes 0325049 (medium-voltage level) and 0325562 (high-voltage level).

References

[1] H. Langkowski, M. Jordan, T.T. Do, D. Schulz, “Spectral Grid Impedance Identification on Different Voltage Levels – Challenges and Realization,” 2017 IEEE PES General Meeting, July 16-20, 2017, Chicago, 2017.

[2] D. Schulz, Power Quality – Theory, Simulation, Measurement and Assessment (in German), VDE-Verlag, Offenbach (2004), ISBN 3-8007-2757-9.

[3] H. Langkowski, M. Jordan, T. Do Thanh and D. Schulz, “Grid impedance determination – identification of neutral line impedance,” 21st International Conference on Electricity Distribution CIRED, Frankfurt, 6-9 June 2011.

[4] H. Langkowski, T. Do Thanh, M. Jordan and D. Schulz, “Grid Impedance Identification considering the Influence of Coupling Impedances,” *IEEE International Symposium on Industrial Electronics (ISIE 2010)*, Bari (Italy), 4-7 July 2010.

[5] M. Jordan, H. Langkowski, T. Do Thanh and D. Schulz, “Frequency Dependent Grid-Impedance Determination with Pulse-Width-Modulation-Signals,” *IEEE 7th International Conference-Workshop Compatibility and Power Electronics CPE 2011*, Tallinn (Estonia), 1-3 June 2011.

[6] K. Heuck, K.-D. Dettmann, and D. Schulz, “Elektrische Energieversorgung”, Vieweg, 9th Edition, Wiesbaden, 2013.

[7] BDEW, “Technical Guideline for Power Generation Plants at Medium-Voltage Grid”, (in German), BDEW, Berlin, 2008.

[8] DIN EN50160, Characteristics of the Voltage in Public Electrical Power Supply networks, (2011).

[9] VDN, Technical Guideline for EEG Power Generation Plants at High and Highest Voltage Grid (in German), VDN, Berlin (2004).

[10] H. Langkowski, T. T. Do, and D. Schulz, “Grid Impedance Determination – Relevancy for Grid Integration of Renewable Energy Systems”, IEEE IECON-09, Porto (2009).

[11] M. Liserre, R. Teodorescu and F. Blaabjerg, “Stability of Photovoltaic and Wind Turbine Grid-Connected Inverters for a Large Set of Grid Impedance Values”, IEEE Transactions on Power Electronics, Vol. 21, No. 1, Jan. 2006.

[12] R. Langella and A. Testa, "A New Method for Statistical Assessment of the System Harmonic Impedance and of the Background Voltage Distortion," *9th International Conference on Probalistic Methods Applied to Power Systems* IEEE, Stockholm, Sweden, June 2006

[13] A. Knop and F. W. Fuchs, "High frequency grid impedance analysis with three-phase converter and FPGA based tolerance band Controller," *Compatibility and Power Electronics (CPE'09),* Badajoz, May 2009

[14] C. Xie, S.B. Tennakoon, R. Langella, D. Gallo, A. Testa, and A. Wixon,“Harmonic impedance measurement of a 25 kV single phase AC supply system,” Proceedings of Ninth International Conference on Harmonics and Quality of Power, Orlando, vol. 1, pp. 214-219, 1-4 Oct. 2000.

[15] M. Nagal, W. Xu, and J. Sawada, “Harmonic impedance measurement using three-phase transients,” IEEE Transactions on Power Delivery, vol. 13, pp. 272-277, January 1998.

[16] M. Jordan, T. Do Thanh, H. Langkowski and D. Schulz, “Strategies for Frequency Dependent Grid Impedance Measurement at the Medium- and High-Voltage Level,” *5th International Ege Energy Symposium and Exhibition (IEESE-5)*, Denizli, 27-30 June 2010.

[17] M. Jordan, F. Grumm, H. Langkowski, T. Do Thanh and D. Schulz, “Online Network Impedance Identification with Wave-Package and Inter-Harmonic Signals,” *XII International School on Nonsinusoidal Currents and Compensation 2015 (IEEE)*, Lagów (Poland), 15-18 June 2015.

[18] H. Langkowski, M. Jordan, T. Do Thanh and D. Schulz, “Grid impedance determination – identification of neutral line impedance,” *21st International Conference on Electricity Distribution CIRED*, Frankfurt, 6-9 June 2011

[19] T. Do Thanh, T. Schostan, T. Dettmann and D. Schulz, “Nonsinusoidal Power Caused by Measurements of Grid Impedances at Unbalanced Grid Voltages,” Proceedings of the IEEE conference of the International School on Nonsinusoidal Currents and Compensation ISNCC, Lagow, 10-13 June 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 15-29 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Design of True Random Numbers Generators with Ternary Physical Unclonable Functions

Bertrand Francis Cambou*

School of Informatics Computing and Cyber Systems, Northern Arizona University, Flagstaff, 86004, USA

ARTICLE INFO

Article history:
Received: 06 April, 2018
Accepted: 02May, 2018
Online: 07 May, 2018

Keywords:
Random numbers generators
Physical unclonable functions
Ternary states

ABSTRACT

Memory based ternary physical unclonable functions contain cells with fuzzy states that are exploited to create multiple sources of physical randomness, and design true random numbers generators. A XOR compiler enhances the randomness of the binary data streams generated with such components, while a modulo-3 addition enhances the randomness of the native ternary data streams, also generated with the same method. Deviations from perfect randomness of these random numbers, in terms of probability to be non-random, was reported as low as 10^{-10} in the experimental section of this paper, which is considered as extremely random based on NIST criteria.

1. Introduction

This paper is an extension of the work presented at the annual computing SAI conference, London, July 2017 [1]. The strengthening of cryptographic protocols with random numbers [2-4] is widely accepted as mandatory to secure networks of cyber physical systems (CPS). Both pseudo random numbers generators (PRNG) that use mathematical methods [5-8] and true random numbers generators (TRNG) that exploit physical elements [9-11] are mainstream. The randomness based on mathematical algorithms for PRNGs could be weak when crypto-analysts armed with powerful computers know the algorithms. Such algorithms can also consume too much computing power, which may be a problem for small internet of things (IoT) peripherals. The need to quantify the randomness of PRNG, and TRNG is of prime importance [12-15]. Physical Unclonable functions (PUF) can be valuable sources of natural randomness [16-17], they have been adopted for the design of true random number generators (TRNG). However, the randomness of the physical elements is not always acceptable when subjected to temperature changes, aging, electromagnetic interferences, and other parametric drifts. The PUFs can be too predictable in some circumstances, which is not necessarily conducive to the design of quality TRNGs that rely on physical randomness to generate a fresh random number at every query.

We are presenting how ternary PUFs contain fuzzy elements that are excellent sources of randomness. The method is based on the identification of the cells of memory based PUFs that are naturally unstable under repetitive queries. When tested, the fuzzy cells can switch back and forth randomly between "1" and "0", thereby generating random data streams. We are presenting three complementary elements:

i) how a XOR data compiler, which process the data available from multiple ternary cells, can create an extremely high level of randomness [18-21];
ii) how a probabilistic model allows the quantification of the level of randomness of the TRNG;
iii) how the method can be extended to the generation of native ternary random numbers with modulo-3 addition.

2. Designing a random number generator

2.1. Ternary physical unclonable functions

There is a growing interest in securing CPS's with Physically Unclonable Functions (PUFs) to strengthen security when deployed in the cryptographic processes using a powerful set of physically derived cryptographic primitives [22-26]. PUFs act as virtual fingerprints for the hardware during the authentication processes to effectively block cyber thefts, Trojans, and malwares [27-34]. With error correcting methods, the PUFs can also generate cryptographic keys for symmetrical encryption schemes [35-36]. The inherent randomness, unclonability, secrecy, and physical nature of most PUFs makes it extremely hard to inspect during side channel attacks, or when lost to the enemy.

2.2. Quality considerations for PUFs

The PUFs, regardless of their design, must exhibit enough predictability overtime for reliable authentications, or encryption. The reference patterns of the PUFs that are

*Bertrand Francis Cambou, Email : Bertrand.cambou@nau.edu

generated up front during the setup of protocol, called challenges, are compared over the life of the component with freshly generated patterns, called responses, during the authentication cycles. Quality PUFs need low challenge-response-pair (CRP) error rates, the intra-PUF challenge-response Hamming distance must be small enough to insure small level of false rejection rates (FRR). The PUF challenges should act as predictable "digital fingerprints" of the component, while the responses should be easily recognizable as a measurement of the same "digital fingerprints". Error rates in the 3-7% range are usually acceptable when combined with error correcting techniques [35-36]. PUFs exploit the device-to-device randomness that is created during the manufacturing process of micro-components; it is desirable that the average inter-PUF hamming distance between different PUFs, divided by the length of the PUFs, should be in the 50% range to insure low level of false acceptance rates (FAR). This is achievable when the level of intra-PUF randomness, also called entropy, is high enough. PUFs with longer streams of bits have therefore higher entropy, and lower FAR. 128-bit CRPs, or higher, are usually required for this purpose.

Another important figure of merit for PUFs is the number of available CRP configurations for a unit. Strong PUFs, as opposed to weak PUFs, contains large quantities of possible CRPs that are addressable. For example, a ring oscillator PUF with 128 rings is a strong PUF. The number of possible pairing of two rings is N $=\binom{128}{2}$ = 16,256. If the protocol use 128-bit long CRPs, the number of possible challenges of 2^{128}, offers satisfactory entropy, and a low collision rate of the pairs. A memory PUF with random addressing capabilities is even stronger [36, 39-40]. For example, when the capacity of the memory is in the mega-byte range, millions of configurations are providing an entropy much higher than a 128-ring oscillator with "only" 16,256 possible configurations. Existing PUFs can have limitations, and lack of trustworthiness that could create a false sense of security. The signatures of PUFs are derived from intrinsic manufacturing variations, which could become predictable due fabrication excursions. Properties such as critical dimensions of printed structures, doping levels of semiconducting layers, and threshold voltages should make each device unique and identifiable from all other devices, abnormal operations during the manufacturing process could alter such randomness. When subject to changes related to temperature, voltage, EMI, aging, and other environmental factors these parameters can drift over time, the undesirable result, is weak PUFs with CRP error rates as high of 20%.

The main objective in designing ternary PUFs is to resolve some of these issues, and to reduce the CRP error rates by eliminating fuzzy CRPs during challenge generation. The figure of merit is to achieve trustworthy and robust intra-PUF CRP matching rates with low FRR during authentications, without increasing FAR during inter-PUF authentication of malicious challenges. The by-product of such design is the design of highly random TRNG with the fuzzy cells.

2.3. Memory based PUFs

The methods to design PUFs and TRNGs with SRAM memories have been published SRAM [36-38]. SRAM based PUFs have been successfully commercialized. When powered up, each SRAM cells naturally flip to store either a 0, or a 1. In most of the cases, arrays of SRAM cells return to a similar pattern characteristic, i.e. a similar finger print. SRAM based PUFs designed with this feature can be reliable, however heavy error correcting methods are usually needed. The SRAM based PUFs are not particularly immune to side channel attacks. Significant research efforts have been published regarding the design of PUFs with Flash RAMs [39-40], DRAMs [41-44], magnetic RAMs [45-46], and resistive RAMs [47-49]]. The cryptographic protocols leveraging memory PUFs are in general distinct from the ones developed with other mainstream PUFs such as ring oscillators, or gate delay arbiters. As shown in Fig. 1, the value of a parameter $\mathcal{P}$ is measured on each cell, and is compared with a threshold. The cells with parameter $\mathcal{P}$ below the threshold are "0"s, and are "1"s above the threshold. Examples of parameter P selected to design memory PUFs include: threshold voltages of Flash cells after fixed time programming; charges left on DRAM cells without refresh; high resistance value of MRAM cells after programming; and Vset of ReRAM cells.

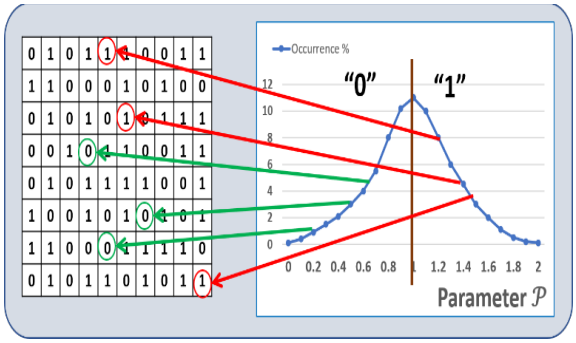


Figure 1: Diagram explaining the design of memory based PUF with a parameter $\mathcal{P}$, and a threshold to sort out the states 0 and 1.

The CRP matching is done after error correction. The cell-to-cell physical parameter variations due to manufacturing variations are too erratic for CRP generation. Fig.2 is a diagram showing how a drift of $\mathcal{P}$ toward the higher value is forcing the responses of the cells located close to the threshold to switch from 0 to 1, which increase CRP error rates. The cells located far from the transition are not impacted.

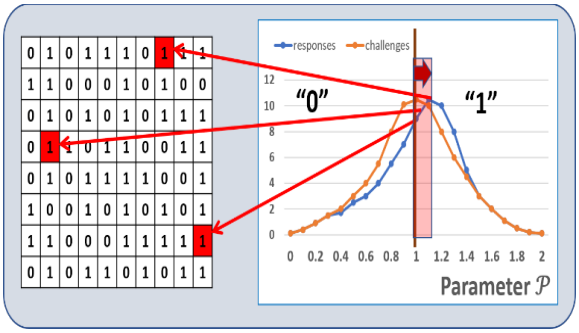


Figure 2: Diagram explaining the impact of a drift of parameter $\mathcal{P}$ toward higher values, creating CRP errors.

2.4. Ternary PUFs

The concept of memory based PUFs with ternary states having random number generation capabilities is described [49-

52]. The measurement of $\mathcal{P}$ of the cells of a memory PUF allows the segmentation of the cell population into three states. The cells with $\mathcal{P} < \mathbf{T_1}$ (a low threshold) carry the state "-", the cells with $\mathcal{P} > \mathbf{T_2}$ (a high threshold) carry the state "+", and the remaining cells carry the ternary state "0" cells, see Fig.3.

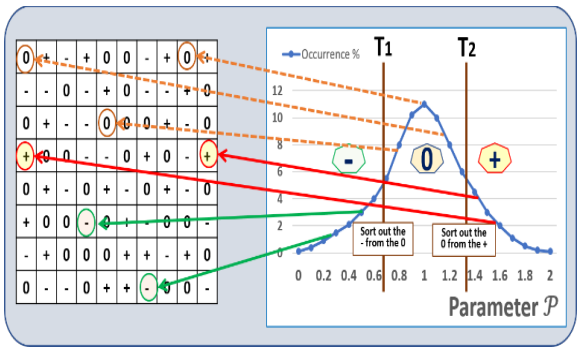


Figure 3: Diagram describing the sorting of the memory PUF into ternary states based on parameter $\mathcal{P}$.

During challenge generation, the cells are sorted into ternary states. During response generation, only the cells with "-" or "+" states are queried, while the cells with "0" state are ignored, see Fig.4. The PUF CRP error rates are significantly lowered, the distance $\mathbf{T_2}$-$\mathbf{T_1}$ acts as a buffer between the states "-", and "+". When the distance $\mathbf{T_2}$-$\mathbf{T_1}$ between thresholds increases, the CRP error rate can reach extremely low values, and is less sensitive to various drifts.

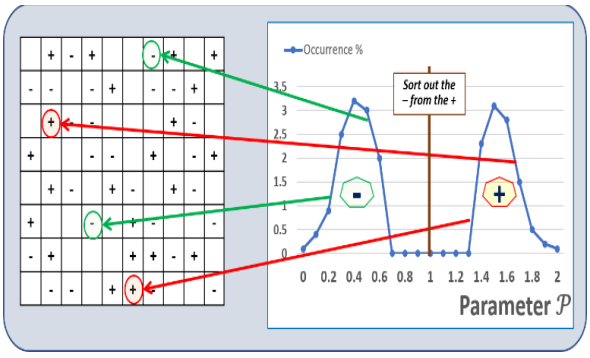


Figure 4: Diagram showing the response generation. The "0" states are ignored, only the cells with "-" and "+" challenges are considered

3. Random number generators

3.1. Pseudo Random number generators

There are numerous excellent PRNG available to the system developers, which are highly reliable [4-8]. For example, a PRNG $\{a_1, a_2, \ldots, a_n\}$ can be designed with congruential generators, where a is the multiplier, c the increment, m the modulus, and X_i , b, c, m are natural numbers, typically, c and m are chosen to be relatively prime:

$$a_{i+1} = (b\, a_i + c) \bmod m \qquad (1)$$

Other example of PRNG can constructed by using iterative encryption, as shown in Fig 5, a_{i+1} is the cipher of Xi which is encrypted by the code E, and the key K_i . Proving that a PRNG or a TRNG is "random" is a very complicated task that could take years to validate, and billions of data points. The National Institute of Standard and Technology (NIST) has developed an excellent suite of tools available on line that can test the randomness of any random numbers generators [12-15].

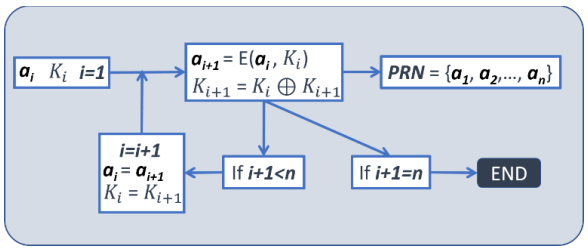


Figure 5: Generating the random number PRN from two numbers a_i K_i, and the encryption scheme E.

Examples of parameters that are tested include deviation from randomness, a frequency test (monobit test), Serial test (two-bit test), a Poker test (non-overlapping parts), run tests (gap and blocks), and autocorrelation tests [Menezes, van Oorschot, Vanstone - Handbook of Applied Cryptography].

In this paper, we are using statistical analysis to quantify randomness, and the parameter λ defined below in this section. Each bit "a_i" of a data stream of n bits should have a 50% probability to be either a "1" or a "0". The average deviation from perfect randomness λ is given by:

$$\mathbf{P}_{(\mathbf{a_i}=1)} = \mathbf{P}_{(\mathbf{a_i}=0)} = 0.5 \qquad (2)$$

$$\lambda = |\, 0.5 - \mathbf{P}_{(\mathbf{a_i}=1)} \,| = |\, 0.5 - \mathbf{P}_{(\mathbf{a_i}=0)} \,| \qquad (3)$$

Assuming that the length of the data stream is n =128, with $\mathbf{P}_{(\mathbf{a_i}=0)} = 0.5$, the number of possible combinations, also called entropy, is $2^{128} = 3.4\ 10^{38}$, which is large enough to protect cryptographic functions from existing or foreseeable computers. When the RNG is not totally random, in this case $\lambda \neq 0$, the entropy is lower than 2^{128}, and is further reduced with larger λ. A position paper from (NIST) [12], suggested in 1999 that λ greater than 10^{-3} would not be acceptable, sophisticated crypto-analysis methods could be effective to break the PRNG. NIST in 2010 and others [33-34] revisited this. The value of λ that is acceptable to get a safe TRNG is a moving target as modern computers get increasingly powerful. To the best of our knowledge, $\lambda < 10^{-5}$ is currently considered an excellent target, while $\lambda < 10^{-10}$ is considered outstanding.

3.2. Use of XOR to enhance PRNGs

Exclusive OR, XOR, is a Boolean logic gate widely adopted in cryptography [18-21]. Two input bits a_i and a_{i+1} are transformed into $c_i = a_i \oplus a_{i+1}$, with the following equations:

$$c_i = 0 \text{ if } a_i = a_{i+1}\ (0 \oplus 0 \text{ or } 1 \oplus 1) \qquad (4)$$

$$c_i = 1 \text{ if } a_i \neq a_{i+1}\ (0 \oplus 1 \text{ or } 0 \oplus 1) \qquad (5)$$

XOR logic is part of most encryption algorithms such as the Data Encryption System (DES), the Advanced Encryption System (AES), and hash functions such as SHA. XOR functions are adding confusion and randomization in the encryption process while been reversible in the decryption process. As part of the encryption, the data streams generated by plain texts are

often XORed with cryptographic keys, or sub-keys, then XORed again during decryption. XOR scramblers can enhance randomization in multicarrier communications [19]. XOR are also used to generate scrambling sequences to achieve data randomization in a memory circuit, as well as enhancing random number generators [20]. Some of the important reasons for the use of XOR functions in cryptography are:

- c_i is not disclosing the value of a_i and a_{i+1}:
  c_i =0 can be the result of the pair 00, or the pair 11;
  c_i =1 can be the result of the pair 01, or the pair 10;
- XOR is a symmetrical function when applied twice:

$$a_i \oplus a_{i+1} \oplus a_{i+1} = a_i \quad (6)$$

- If two bits a_i and a_{i+1} are random, the bit c_i, defined by $c_i = a_i \oplus a_{i+1}$, is even more random than a_i or a_{i+1} .

These properties are exploited in the design of the XOR data compiler as presented in section 4.

3.3. Ternary PUFs as sources of randomness

The cells of a ternary PUF with "0" state, as described in section 2.3, are exploited as sources of randomness to design TRNG [1, 17], as explained in Figure 6. The cells located in the center of the distribution, the "0" states, can flip back and forward when the value of their parameter $\mathcal{P}$ is compared to a threshold centered in the median point of the distribution.

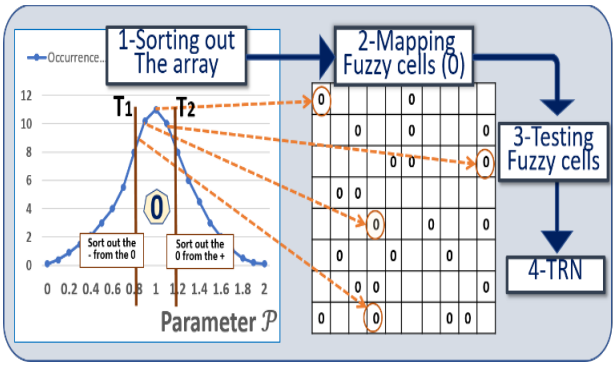


Figure 6: Diagram showing how the cells with "0" states can be sources of randomness

When the distance $\mathbf{T_2}$-$\mathbf{T_1}$ between the two thresholds used to select the cells with "0" states is reduced, the probability to test these cells either as below the median, or above the median at each query is closer to 50%. For example, the selection of 1,000 cells located close to the median will represents a strong pool to design TRNG. These 1,000 cells can be queried many times to generate long random numbers. Each cell acts as a single source of independant randomness subject to noise, and measurement uncertainties. Within the cells of a particular memory array, the distribution of the physical parameter $\mathcal{P}$, which determine if a cell is a "0" or a "1", is following a distribution with a standard variation $\boldsymbol{\sigma_{Array}}$ due to cell-to-cell variations created during manufacturing, and other instabilities. Repetitive measurements of parameter $\mathcal{P}$ on the same cell follow a distribution with the standard variation $\boldsymbol{\sigma_{Cell}}$ responding to various measurement instabilities, noise, and environmental variations. Low error rate PUFs, with predictable CRPs, should have these variations verifying:

$$\boldsymbol{\sigma_{Cell}} \lll \boldsymbol{\sigma_{Array}} \quad (7)$$

When the variations within cells are much lower than the cell-to-cell variations, the "finger print" of the memory PUF is stable and predictable. On the opposite side, to design a TRNG, it is desirable to select only the cells extremely close to the transition of parameter $\mathcal{P}$ between "0' and "1", i.e. the one with ternary state "0". If $\mathbf{T_M}$ is the median of the distribution, the average value $\mathbf{Tx}$ of $\mathcal{P}$ of each -cells should be such that:

$$|\mathbf{Tx} - \mathbf{T_M}| \lll \boldsymbol{\sigma_{Cell}} \quad (8)$$

This maximizes the chance of a random number to flip between "0s" and "1s". In order to enhance the level of randomness only a very small percentage of the memory arrays are selected as sources of randomness. Current secure micro-controllers have very large embedded memory density, typically in the 1 to 100 Mbits, the percentage of the array consumed for TRNG can be relatively small. In the following sections, we are developing a statistical model to study how to enhance the randomness of a data stream generated from the fuzzy cells. One of the tradeoffs to model is the compromise between tightly selecting the "0" cells around the median $\mathbf{T_M}$, versus improving randomness; in the case of the generation of native ternary streams, we study the use of modulo3 adders.

4. Modeling a ternary PUF for TRNG

As shown in Fig 7, the cells that are sorted as unstable with a "0" state can be segmented into two subgroups:

- The cells that have a higher probability to be tested above the median are called A-cells, see Fig 8. They have an higher average probability $\mathbf{P_A}$ to generate a "1"in the stream of random numbers, their average deviation to randomness is $\boldsymbol{\lambda_A}$. The A cells have an average probability $\boldsymbol{P'_A}$ to generate a "0" in the stream of random number:

$$\mathbf{P_A} = 0.5 + \boldsymbol{\lambda_A} \quad (9)$$

$$\mathbf{P'_A} = 0.5 - \boldsymbol{\lambda_A}; \qquad 1 = \mathbf{P'_A} + \mathbf{P_A} \quad (10)$$

- The cells that have a higher probability to be tested below the median are called B-cells, see Fig 9. They have an average probability $\mathbf{P'_B}$ to generate a "0", and an average deviation to randomness $\boldsymbol{\lambda_B}$. The average probability $\boldsymbol{P_B}$ to be generate a "1"is:

$$\mathbf{P'_B} = 0.5 + \boldsymbol{\lambda_B} \quad (11)$$

$$\mathbf{P_B} = 0.5 - \boldsymbol{\lambda_B}; \qquad 1 = \mathbf{P'_B} + \mathbf{P_B} \quad (12)$$

The selection of the transition T_M of parameter $\mathcal{P}$ can be such that the number of A-cells equal the number of B-cells and:

$$\mathbf{P_A} = \mathbf{P'_B}, \quad \mathbf{P'_A} = \mathbf{P_B}, \quad \boldsymbol{\lambda_A} = \boldsymbol{\lambda_B}. \quad (13)$$

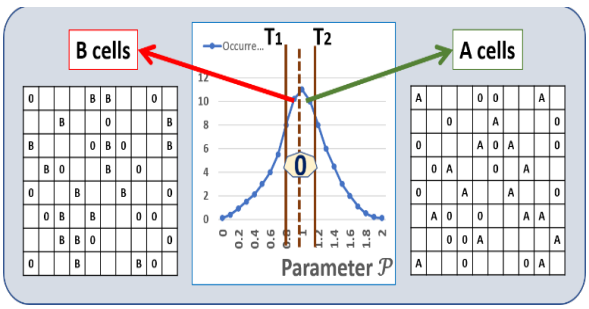


Figure 7: The "0" states are segmented into the A cells that more often measured above the median, and the B cells below the median.

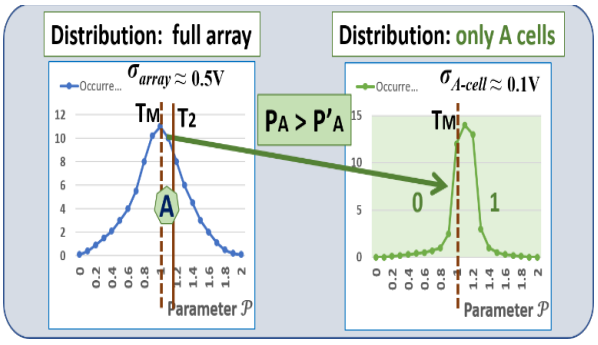


Figure. 8: A-cells with higher probability P_A to generate a 1.

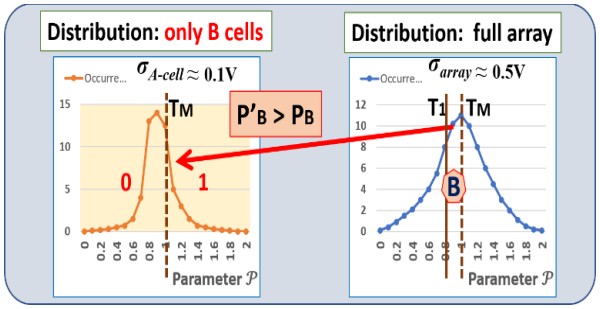


Figure. 9: B-cells with higher probability P'_B to generate a 0.

5. A XOR data compiler for TRNG

As presented below in the experimental section, with 2% of the cell population selected as fuzzy 0-cell, $\lambda_A = \lambda_B \approx 2\ 10^{-2}$, which is far from the level of randomness needed to generate quality TRNG, this based on NIST criteria. In this section, a XOR compiler is developed, with the objective to enhance the level of randomness of the resulting streams, see Fig.10.

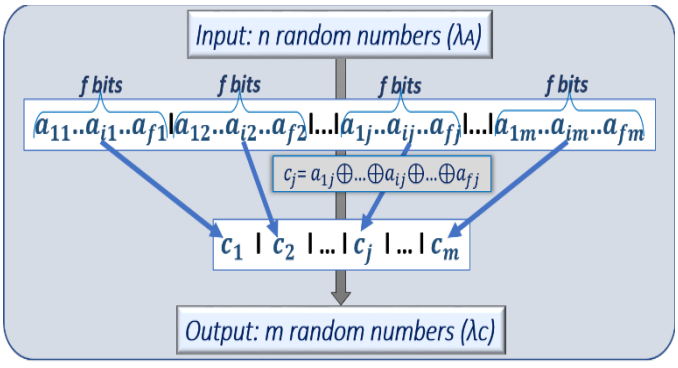


Figure. 10: Description of the effect of the XOR operations

The XOR compiler transforms the incoming streams a_i, $i \in \{1 \text{ to } n\}$, generated by the memory PUF by out coming streams c_j, $j \in \{1 \text{ to } m\}$, $m<n$, of higher level of randomness. The stream of n random numbers generated from the ternary memory PUF is shown below.

$$\textit{Incoming stream:} \qquad \{a_1, a_2, \ldots, a_i, \ldots a_n\} \qquad (14)$$

This stream is grouped in chunks of f bits, $i \in \{1 \text{ to } f\}$; $f < n$.

$$\textit{Chunk of bits:} \qquad \{a_{1j}, a_{2j}, \ldots, a_{ij}, \ldots, a_{fj}\} \qquad (15)$$

For example, 1,280 random bits are grouped in 128 chunks of 10 bits. With a XOR, the stream c_j, with $j \in \{1 \text{ to } m\}$ and $n=m.f$, is generated from the stream a_i, as shown in Fig 10.

$$\textit{Out coming stream:} \qquad \{c_1, c_2, \ldots, c_j, \ldots, c_m\} \qquad (16)$$

$$c_j = a_{1j} \oplus a_{2j} \oplus \ldots \oplus a_{ij} \oplus \ldots \oplus a_{fj} \qquad (17)$$

Such a XOR compiler can be implemented in hardware with only a few logic gates which can be inserted as part of the crypto-processor of the secure processor. The PRNGs presented section 3.1, are generated sequentially, the random number a_{i+1} of a stream of n bits is generated from the previous random numbers a_i. Conversely, the TRNGs with XOR gates can be generated in parallel eq (17) in one cycle. XOR gate is also an addition modulo 2 without carry over. A quicker way to compute eq (17) is to count how many "1s" are presents in the stream $\{a_{1j}, a_{2j}, .., a_{fj}\}$. If the number of "1s" in the stream is odd then c_j =1, when even c_j=0.

$$c_j = a_{1j} \oplus a_{2j} \oplus \ldots \oplus a_{fj} = (a_{1j} + a_{2j} + \ldots + a_{fj}) \bmod 2 \qquad (18)$$

6. Modeling a 2-bit XOR compiler

We analyze a 2-bit XOR compiler, the incoming data stream of $2n$ bits is "XORed" two bits by two bits to generate a stream of n bits, f=2. There are three possible configurations for each "XORing": both cells are A-cells, one cell is an A-cell and the second is a B-cell, and both cells are B-cells. Let us choose: $P_A = P'_B = 0.52$; $P'_A = P_B = 0.48$; $\lambda_B = \lambda_A = 2\ 10^{-2}$.

➢ *Number of A-cells is even: two A-cells, or two B-cells*

The probability P'_C to have $c_j = a_{1j} \oplus a_{2j}$ at "0" is occurring when the two cells (a_{1j}, a_{2j}) are at (00) or (11):

$$P'_C = P'^2_A + P^2_A = 0.5008 \rightarrow \lambda_C = 8\ 10^{-4} \qquad (19)$$

The probability P_C to have $c_j = a_{1j} \oplus a_{2j}$ at "1" is occurring when the two cells (a_{1j}, a_{2j}) are at (01) or (10):

$$P_C = 2\ (P_A\ P'_A) = 0.4992 \qquad (20)$$

➢ *Number of A-cells is odd: one A-cells, and one B-cell.*

The probability P_C to have $c_j = a_{1j} \oplus a_{2j}$ at "1" is occurring when the two cells (a_{1j}, a_{2j}) are at (01) or (10):

$$P'_C = P^2_A + P'^2_A = 0.5008 \rightarrow \lambda_C = 8\ 10^{-4} \qquad (21)$$

The probability P'_C to have $c_j = a_{1j} \oplus a_{2j}$ at "0" is occurring when the two cells (a_{1j}, a_{2j}) are at (00) or (11):

$$P_C = 2\ (P_A\ P'_A) = 0.4992 \qquad (22)$$

Let us assume that the incoming stream with $2n$ bits is generated from a memory PUF with 50% A-cells and 50% B-cells, and with $\lambda_A = 2\ 10^{-2}$. The 2-bit XOR compiler can statistically generate an out coming stream of n bits having 50% C-cells, and 50% D-cells with $\lambda_C = 8\ 10^{-4}$, see Fig. 11. The C-cells are made of pairs of either AA cells or BB cells, while the D-cells are made of pairs of either AB cells or BA cells. In both cases, f =2 is even. The general equations developed below in section 8, eq. (29) to (39) are applicable. When the number of B-cell is even $P_C < P'_C$, and are reversed when the number of B-

cell is odd $\mathbf{P_C} > \mathbf{P'_C}$. The deviation from randomness λ_C=8 10^{-4} is 25 times smaller than the deviation before the 3bit-XOR compilation, λ_A=2 10^{-2}.

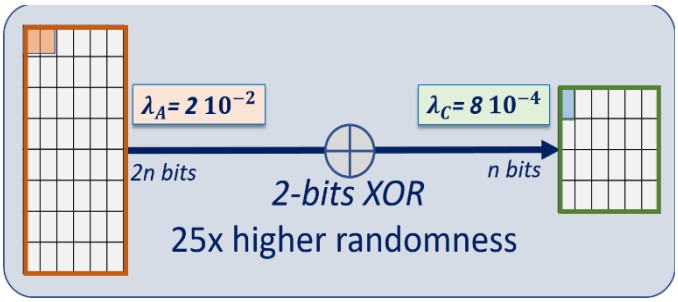


Figure. 11: Diagram showing a 2-bit XOR compiler.

7. Modeling a 3-bit XOR compiler

In this section we analyze a 3-bit XOR compiler, in which the incoming data stream of 3***n*** bits is "XORed" three bits by three bits to generate a stream of ***n*** bits, ***f***=3. We are again choosing the same example:

$\mathbf{P_A}=\mathbf{P'_B}$=0.52; $\mathbf{P'_A}=\mathbf{P_B}$=0.48; $\lambda_B=\lambda_A$=2 10^{-2}.

There are four possible configurations for each "XORing": three cells are A-cells, two cells are A-cells & one cell is B-cell, one cell is A-cell & two cells are B-cells, and finaly three cells are B-cells.

- *Number of B-cells is even: three A-cells, A-cell & two B-cells.* The probability $\mathbf{P_C}$ to have $c_j=a_{1j}\oplus a_{2j}\oplus a_{3j}$ at "1" is occurring when the three cells (a_{1j}, a_{2j}, a_{3j}) are at (111), (100), (010) or (001):

$$\mathbf{P_C} = \mathbf{P_A^3} + 3\ \mathbf{P_A P'^2_A} = 0.500032 \rightarrow \lambda_C = 3.2\ 10^{-5} \qquad (23)$$

The probability $\mathbf{P'_C}$ to have $c_j=a_{1j}\oplus a_{2j}\oplus a_{3j}$ at "0" is occurring when the three cells (a_{1j}, a_{2j}, a_{3j}) are at (000), (110), (011) or (101):

$$\mathbf{P'_C} = \mathbf{P'^3_A} + 3\ \mathbf{P'_A P^2_A} = 0.499968 \qquad (24)$$

- *Number of B-cells is odd: Three B-cells, B-cell & two A-cells.* The probability $\mathbf{P'_C}$ to have $c_j=a_{1j}\oplus a_{2j}\oplus a_{3j}$ at "0" is occurring when the three cells (a_{1j}, a_{2j}, a_{3j}) are at (000), (110), (011) or (101):

$$\mathbf{P'_C} = \mathbf{P_A^3} + 3\ \mathbf{P_A P'^2_A} = 0.500032 \rightarrow \lambda_C = 3.2\ 10^{-5} \qquad (25)$$

The probability $\mathbf{P'_C}$ to have $c_j=a_{1j}\oplus a_{2j}\oplus a_{3j}$ at "0" is occurring when the three cells (a_{1j}, a_{2j}, a_{3j}) are at (000), (110), (011) or (101):

$$\mathbf{P_C} = \mathbf{P'^3_A} + 3\ \mathbf{P'_A P^2_A} = 0.499968 \qquad (26)$$

Let us assume that the incoming stream with 3***n*** bits is generated by a memory PUF having 50% A-cells, and 50% B-cells, and with λ_A=2 10^{-2}. The 3-bit XOR compiler can statistically generate an out coming stream of 128 bits having 50% C-cells, and 50% D-cells with λ_C=3.2 10^{-5}, see Fig. 12. The C-cells are made of triplets of either AAA cells, ABB cells BAB cells, or BBA cells. The D-cells are made of triplets of either AAB cells, ABA cells, BAA cells, or BBB cells.

In both cases, ***f***=3 is odd. The general equations developed in the next section, eq. (29) to (38) are applicable. When the number of B-cells is even, $\mathbf{P_C} < \mathbf{P'_C}$, and are reversed when the number of B-cells is odd, $\mathbf{P_C} > \mathbf{P'_C}$. The resulting deviation from randomness, λ_c=3.2 10^{-5}, is 25x25=625 times smaller than the deviation before the 3-XOR compilation, λ_A=2 10^{-2}. It is interesting to notice that a 3-bits XOR data compiler needs only 50% more starting cells than a 2-bits compiler, and has a level of non-randomness 25 times lower.

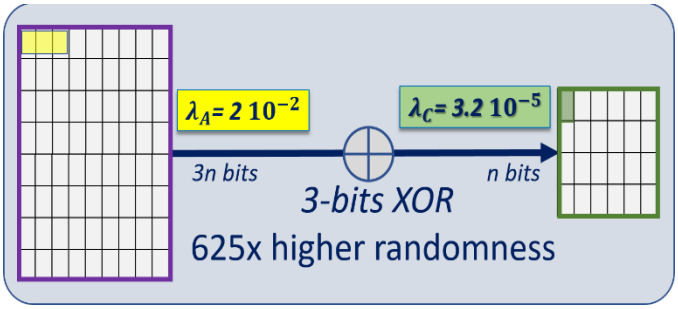


Figure. 12: Diagram showing a 3-bits XOR compiler.

8. Modeling the XOR compiler in general terms

The goal is to develop a model that quantifies the effect of a XOR compiler, which enhance the level of randomization of a data stream, as a function of the size ***f*** of the chunk of incoming bits that are XORed together. The incoming stream {a_1, …, a_i, …a_n} has a deviation from randomness λ_A , and the out coming stream {c_1, c_2, …, c_j, …, c_m} has a deviation λ_C. This variation is obtained by computing P_C, the probability for c_j, to be a "1:

$$|\ \lambda_C\ | = 1 - \mathbf{P_C} \qquad (27)$$

The incoming random bits a_{ij} are generated from A-cells or B-cells of the ternary PUFs. As stated in section 4.1, the transition T_M is selected in such a way that the probability to have an A-cell, and a B-cell is equal to 0.5. If each of the ***f*** long chunks have ***s*** A-cells and ***t*** B-cells with ***s***+***t***=***f***. The numbers of possible combinations ***(f, s)*** is:

$$\mathbf{C_{f,s}} = \binom{f}{s} = f!\ /\ s!\ (f\text{-}s)! \qquad (28):$$

1.1 All cells of chunk j are A-cells.

The probability of any of the A-cells of the stream to be a "1" is $\mathbf{P_A}$, and the probability to be a "0"is $\mathbf{P'_A}$. Both $\mathbf{P_A}$ and $\mathbf{P'_A}$ are following Bernoulli formula:

$$1 = \sum_{i=0}^{i=f} \binom{f}{i} \mathbf{P_A^i P'^{f-i}_A} \qquad (29)$$

$$1 = \sum_{i=0}^{i=f} [i \bmod 2] \binom{f}{i} \mathbf{P_A^i P'^{f-i}_A} + \sum_{i=0}^{i=f} [i+1 \bmod 2] \binom{f}{i} \mathbf{P_A^i P'^{f-i}_A}$$

$$1 = \quad P_c \quad + \quad P'_c \qquad (30)$$

The terms $\mathbf{P_A^i P'^{f-i}_A}$ of eq. (29) and (30) correspond to a configuration where ***i*** bits are "1s", and ***f-i*** bits are "0".

<u>The probability $\mathbf{P_C}$, is the sum of all terms having ***i*** odd:</u>

$$i \bmod 2 = 1 \qquad i+1 \bmod 2 = 0 \qquad (31)$$

$$\mathbf{P_c} = \sum_{i=0}^{i=f} [i \bmod 2] \binom{f}{i} \mathbf{P_A^i P'^{f-i}_A} \qquad (32)$$

If f=2k is even: $\mathbf{P_C}= \sum_{i=0}^{i=k-1} \binom{2k}{2i+1} \mathbf{P_A^{2i+1} P'^{2k-2i-1}_A}$ (33)

If f=2k+1 is odd: $\mathbf{P_C}= \sum_{i=0}^{i=k} \binom{2k+1}{2i+1} \mathbf{P_A^{2i+1} P'^{2k-2i}_A}$ (34)

The probability $\mathbf{P'_C}$ is the sum of all terms having i even:

i mod2 = 0 i+1 mod2 = 1 (35)

$$\mathbf{P'_c} = \sum_{i=0}^{i=f}[(i+1) \bmod 2]\ \mathbf{C_{f,i}}\ \mathbf{P_A^i P'^{f-i}_A} \quad (36)$$

If f=2k is even: $\mathbf{P'_C}= \sum_{i=0}^{i=k} \binom{2k}{2i} \mathbf{P_A^{2i} P'^{2k-2i}_A}$ (37)

If f=2k+1 is odd: $\mathbf{P'_C}= \sum_{i=0}^{i=k} \binom{2k+1}{2i} \mathbf{P_A^{2i} P'^{2k+1-2i}_A}$ (38)

When f is even, $\mathbf{P_C < P'_C}$ is written as $\mathbf{P_C}=0.5-\boldsymbol{\lambda c_j}$ or $\mathbf{P'_C}=0.5+\boldsymbol{\lambda c_j}$, with $\boldsymbol{\lambda c_j}$ the deviation from randomness of c_j.

When f is odd, $\mathbf{P_C > P'_C}$ and is written as $\mathbf{P_C}=0.5+\boldsymbol{\lambda c_j}$ or $\mathbf{P'_C}=0.5-\boldsymbol{\lambda c_j}$.

1.2 Chunks j are a combination of A-cells & B-cells

The f cells randomly contain A-cells and B-cells. The symmetry between the A-cell and the B-cells ($\mathbf{P_A=P'_B}$ and $\mathbf{P'_A=P_B}$) results in the following property:

- If the chunk of bits $\{a_{j1}, a_{j2}, \ldots, a_{jf}\}$ is generated by an even number of B-cells, the probabilities $\mathbf{P_C}$ and $\mathbf{P'_C}$ are the same as if the chunk was only generated by A-cells. If f is even, **Pc** and $\mathbf{P'_C}$ are respectively computed with eq. (33) and (37); if f is odd, **Pc** and $\mathbf{P'_C}$ are computed with eq. (34) and (38).
- If the chunk of bits $\{a_{j1}, a_{j2}, \ldots, a_{jf}\}$ is generated by an odd number of B-cells, the probabilities P_C and P'_C are the opposite of the ones generated by A-cells as described by eq. (37) (33) and eq (38) (34):

If f is even, **Pc** is (eq.(37)): $\mathbf{P_C}=\sum_{i=0}^{i=k} \binom{2k}{2i} \mathbf{P_A^{2i} P'^{2k-2i}_A}$, and $\mathbf{P'c}$ is (eq.(34)): $\mathbf{P'_C}= \sum_{i=0}^{i=k-1} \binom{2k}{2i+1} \mathbf{P_A^{2i+1} P'^{2k-2i-1}_A}$

If f is odd, **Pc** is (eq.(38)): $\mathbf{P_C}=\sum_{i=0}^{i=k} \binom{2k+1}{2i} \mathbf{P_A^{2i} P'^{2k+1-2i}_A}$, and $\mathbf{P'c}$ is (eq.(35)): $\mathbf{P'_C}=\sum_{i=0}^{i=k} \binom{2k+1}{2i+1} \mathbf{P_A^{2i+1} P'^{2k-2i}_A}$

1.3 Simplification of the model

The objective of this model is to calculate the absolute deviation from perfect randomness, it is not important to know if $\mathbf{P_C > P'_C}$, or if $\mathbf{P'_C > P_C}$. In all cases, $|\boldsymbol{\lambda c_j}|$ is the statistical deviation from pure randomness, regardless of $\mathbf{P_C}$ being greater or lower than $\mathbf{P'_C}$. Therefore, assuming that all cells are A-cells is simplifying the computation without reducing the accuracy of the model.

9. Experimental analysis with XOR compiler

9.1. Variations of ReRAM memory PUFs

The experimental data presented in this paper is based on the study of resistive random-access memory (ReRAM). The cells of ReRAMs, see references [53-60], are constructed with stacks of two electrodes separated by solid electrolytes, the first one is active to REDOX cycles, and the second one is inert. As shown in Figure 13, differential voltages applied on these stacks can move positively, or negatively, elements such as positive oxygen vacancies or positive metallic cations, which result in varying the resistance of the stacks. The basic physical effect described in Fig. 13, can be achieved with several manufacturing technologies:

- Conductive bridge random access memories (CBRAM) that are based on the conduction of cations such as Ag+, or Cu+ through solid chalcogenide electrolytes, or porous silicon [53-58]. The active electrodes could be made of copper, or silver, while the inert electrode can be fabricated with tungsten;
- Memristors devices can operate as ReRAM, or act as active Boolean gates [59-60]. The conductive filaments usually contain oxygen vacancies. The solid electrolyte can be fabricated with HfO, or TaOx.

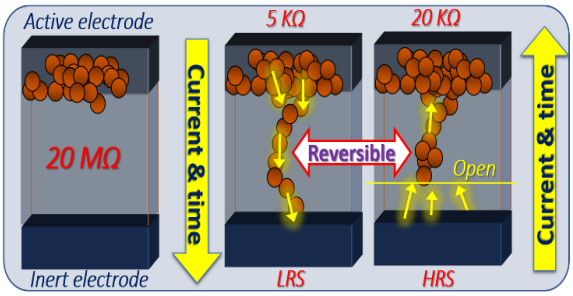


Figure. 13: Diagram showing the programming-erase cycles of a ReRAM. After initial forming, the operations are reversible.

In this work, we had access to Cu/TaOx/Pt resistive crossbar arrays fabricated on thermally oxidized Si wafers, Reference [38]. The Cu/TaOx/Pt switches from "0" to "1" based on the formation and the rupture of filaments, made of oxygen vacancies, bridging the dielectric between both electrodes. The initial conditioning of the ReRAM cells, in which conductive filaments are formed, typically requires a positive voltage of approximately 2 to 5 Volt. After forming, the cells can respond to programming and erasing cycles. It exists a minimum negative Vreset voltage applied across the cells, in the -0.5 to -3.0 Volt range, that force the positive ions or oxygen vacancies to migrate back, breaking the conductive filament. The resulting high resistance state (HRS) is then in the 20 Mohm range. In the positive direction, a minimum Vset voltage applied across the switch, reposition the positive ions or oxygen vacancies, forming again the conductive filament. As shown in Fig. 14, when the voltage is ramping, the current remains low until Vset is reached, then the current quickly increases. This effect is reversible, and the filaments can partially be dissolved with opposite voltages.

The parameter $\mathcal{P}$, that is analyzed for the purpose of designing TRNGs, is the distribution of the Vset across the cells of ReRAM arrays. The entire population of all cells of the array has a Vset distribution that is well represented by a normal distribution having a standard variation $\boldsymbol{\sigma_{Array}}$=0.5V and a median value of 2.1V . The repetitive measurement of the Vset of each cells is also well represented by a normal distribution having a standard distribution $\boldsymbol{\sigma_{Cell}}$= 0.1V.

For the purpose of random number generation, the Vset of each cell is measured; a cell is considered as a "0" state when

Vset<2.1V, and a "1" state when Vset>2.1V. The cells having average Vset measurements at or close to 2.1 Volt, are good candidates for TRNG. The five populations described below are subsets of the total distribution of cells of the array:

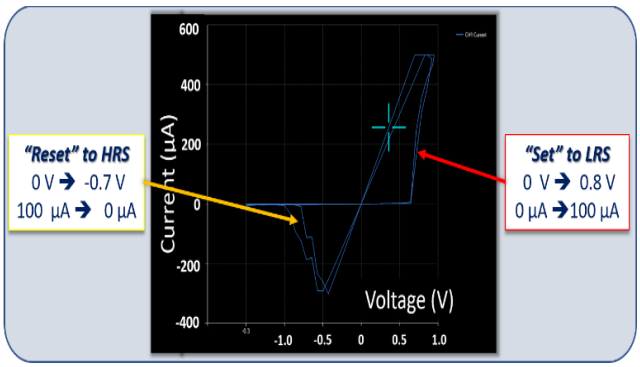


Figure. 14: Experimental characterization of the programming-erase cycles of a ReRAM. The cells are responding to positive, and negative voltage ramps, showing Vset, and Vreset.

- Case-1: Only 2% of the cells are the ternary 0-states. They are used to generate the random numbers. For these cells parameter $\mathcal{P}$ is close to the transition of 2.1 Volt. Half of the cells, the A-cells, have $\mathbf{P_A}$=0.52 probabilities to be "1", $\mathbf{P'_A}$=0.48 probabilities to be "0", with $\lambda_A=2\ 10^{-2}$. The second group, the B-cells, have $\mathbf{P_B}$=0.48 probabilities to be "1", $\mathbf{P'_B}$=0.52 probabilities to be "0", with $\lambda_B=\lambda_A=2\ 10^{-2}$.

$$\mathbf{P_A=P'_B}=0.52;\ \mathbf{P'_A=P_B}=0.48;\ \lambda_B=\lambda_A=2\ 10^{-2} \quad (39)$$

- Case-2: **4%** of the cells are ternary 0-states. The probabilities as defined above are:

$$\mathbf{P_A=P'_B}=0.54;\ \mathbf{P'_A=P_B}=0.46;\ \lambda_B=\lambda_A=4\ 10^{-2} \quad (40)$$

- Case-3: **7%** of the cells are ternary states. The probabilities as defined above are:

$$\mathbf{P_A=P'_B}=0.56;\ \mathbf{P'_A=P_B}=0.44;\ \lambda_B=\lambda_A=6\ 10^{-2} \quad (41)$$

- Case-4: **11%** of the cells are ternary states. The probabilities as defined above are:

$$\mathbf{P_A=P'_B}=0.60;\ \mathbf{P'_A=P_B}=0.40;\ \lambda_B=\lambda_A=1\ 10^{-1} \quad (42)$$

- Case-5: **100%** of the cells are included. The probabilities as defined above are:

$$\mathbf{P_A=P'_B}=0.90;\ \mathbf{P'_A=P_B}=0.10;\ \lambda_B=\lambda_A=4\ 10^{-1} \quad (43)$$

In this last case, there are no ternary states, the entire memory array is used to generate random numbers.The reason we are considering this range of options is to quantify the effectiveness of the XOR data compiler to generate a random number as a function of how tight the ternary state distribution is. Case-1 is the one with the highest initial randomness, while Case-5 is the lowest one.

9.2. Effect of the XOR compiler on the TRNG

The probabilistic model presented in this section is used to analyze the five experimental cases presented above. Fig 15. and Fig. 16 summarize the impact of the XOR data compiler when $\boldsymbol{f}$ varies from 2 to 5. We are observing a lack of efficiency of the XOR compiler in case-5, the one without ternary states. The lack of initial randomness of this case is such that the XORing cannot "clean up" the stream. In other cases, the XOR data compiler when combined with the ternary 0-states is very efficient. Case-1 with the highest level of initial randomness is benefiting the most from the XOR compiler: with 5-cell XOR, $\lambda_C=5.12\ 10^{-8}$, ***which*** is a very small deviation from absolute randomness.

Case	% of 0-cells	λ_A Initial	λ_C 2-bit XOR	λ_C 3-bit XOR	λ_C 4-bit XOR	λ_C 5-bit XOR
1: 52% - 48%	2%	$2\ 10^{-2}$	$8\ 10^{-4}$	$3.2\ 10^{-5}$	$\mathbf{1.28\ 10^{-6}}$	$\mathbf{5.12\ 10^{-8}}$
2: 54% - 46%	4%	$4\ 10^{-2}$	$3.2\ 10^{-3}$	$2.56\ 10^{-4}$	$2.15\ 10^{-5}$	$\mathbf{1.64\ 10^{-6}}$
3: 56% - 44%	7%	$6\ 10^{-2}$	$7.2\ 10^{-3}$	$8.64\ 10^{-4}$	$1.04\ 10^{-4}$	$1.2\ 10^{-5}$
4: 60% - 40%	11%	$1\ 10^{-1}$	$2\ 10^{-2}$	$4\ 10^{-3}$	$8\ 10^{-4}$	$1.6\ 10^{-4}$
5: 90% -10%	100%	$4\ 10^{-1}$	$3.2\ 10^{-1}$	$2.38\ 10^{-1}$	$2.05\ 10^{-1}$	$1.64\ 10^{-1}$

Figure 15: Deviation from non-randomness by experimental case

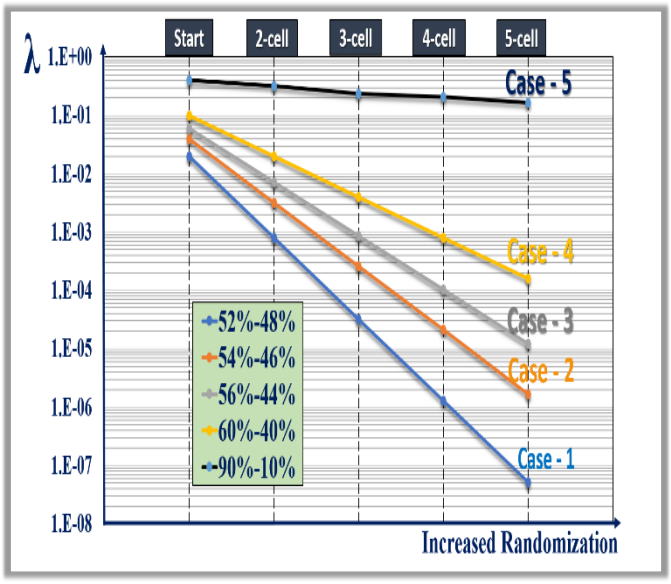


Figure 16: Increased efficiency of the XOR compiler

10. Example of algorithms for TRNG

10.1. Minimization of the impact of parameter drifts

The randomness of the TRNG originates from the physical parameters of multiple cells that provide independent sources of physical randomness. This is a fundamental strength compared with mathematically generated pseudo RNG (PRNG) because mathematical algorithms cannot describe unclonable physical elements. The cell-to-cell randomness is due to micro-variations during manufacturing and natural noise effect during measurements. However, physical elements can vary often in a predictable way when subject to effects such as temperature change, biasing conditions, and induced attacks. For example, the value of the Vset of a resistive RAM goes down when subject to higher temperature. A hacker could submit the physical element to a hot air blower to increase temperature,

reduce Vset, thereby making both A-cells and B-cells appear similar, creating a high probability to be tested as "0". Such a drift or malicious attack could results in a collapse of the level of randomness with lower entropy. The remedy of such an attack is to make the size of the population of A-cells and B-cells equivalent, by adjusting the threshold (T2) between the "0" states, and "1" states, in the median point of the 0-cell distribution, see Fig 17:

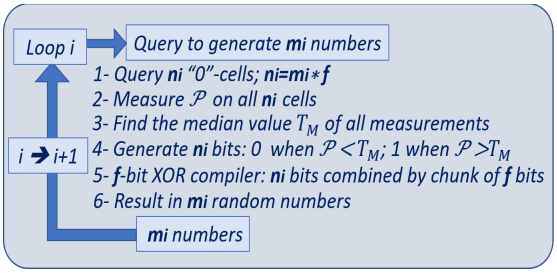


Figure. 17: Algorithm to reduce the effect of a parameter drift.

1) Identify n_i cells of the memory PUF that are part of the fuzzy 0-cells; $n_i = m_i * f$, in preparation of the f-bit XOR compiler;
2) Measure parameter $\mathcal{P}$ of all these ni cells;
3) Identification of the threshold $\mathbf{T_M}$ placed at the median value of all measurements of parameter $\mathcal{P}$ of the population. By design, half of the cells should have a value below $\mathbf{T_M}$, and half above $\mathbf{T_M}$.
4) Generate n_i bits, "0"s below $\mathbf{T_M}$, and "1"s above $\mathbf{T_M}$.
5) Use the XOR compiler to combine chunks of f bits together.
6) The resulting stream of m_i bits is the stream of the TRNG.

With this method the raw data stream generated by the memory array and the 0-cell has a population with equal numbers of "0"s and "1", regardless of a potential drift in temperature caused by a natural variation, or caused by the hot air blower of the hacker. The method is applicable to compensate for any drifts, noise, or aging; the integrity of the TRNG is thereby protected.

10.2. Generalization to other TRNG designs

When the physical component generates a data stream with a deviation from absolute randomness λ_{in}, it is possible to model the size f of the chunks that are XORed together, as described in Fig.8, to meet a particular λ_{out} objective. The model can be used as a predictive tool. For example, as shown Fig.18, the number f necessary to compile a data stream of various initial randomness can be anticipated to be $\lambda_C < 5\ 10^{-8}$ or $\lambda_C < 10^{-10}$. This could be valuable to adjust the compilation as a function of the monitoring of the randomness of the incoming data stream.

Initial Randomness	% of 0-cells	For λ_C<5 $\mathbf{10^{-8}}$ *f-bit* XOR	For λ_C<1 $\mathbf{10^{-10}}$ *f-bit* XOR
52% - 48%	2%	5	7
54% - 46%	4%	7	9
56% - 44%	7%	8	12
60% - 40%	11%	10	14

Figure 18: Predictive model – f-bit needed for a given objective λ

The proposed method to design TRNG is not limited to ReRAM arrays, and Vset as parameter $\mathcal{P}$. The method is applicable to any memory device as long as it is possible to identify a parameter $\mathcal{P}$ that can be reliably tested to sort out the cells and identify enough unstable 0-cells. The algorithm presented Fig. 17 is generic:

- Flash or EEPROM memory: parameter $\mathcal{P}$ can be the trans-conductance of the cells after fixed time programming. The threshold voltage of each cell, after fixed time injection of electrons in the floating gate, vary cell-to-cell due to variations in fabrication parameters such as tunnel oxide thickness and doping levels. Very small changes of threshold voltage can create major changes in the trans-conductance, which are desirable sources of randomness.
- DRAM memory: parameter $\mathcal{P}$ could be the measurement of the residual charge left in a cell after constant discharging time. One effective method is to program all cells, and put the refresh cycle on hold. The fuzzy cells can flip above or below the threshold value of residual charge.
- ReRAM memory: In addition of the Vset as presented in this paper, parameter $\mathcal{P}$ could be the Vreset (threshold voltage to erase the cells), Roff (resistivity on the high resistance state), or Ron (resistivity on the low resistivity state). Some parameters like Roff can be flaky, and jump in a non-erratic way from a set of several discrete values, which is not a desirable source of randomness for a TRNG.
- SRAM memory: the PUFs are based on the determinations of the cells flipping to either a 0 state, or a 1 state after power-off- power-on cycles. However, 3 to 5% of the SRAM cells are fuzzy, they can switch on either states at each cycle. The recommended methodology is to test the SRAM array, and keep track of the 0-states for TRNG.

The use of the XOR compiler enhances the randomness of any data streams regardless of their origin. The XORing by chunk of f-bits is therefore applicable to a stream of n_i incoming bits, as shown in Fig.19.

1) If the length of the incoming stream of n_i bits is not an integer number multiple of f, the two are related by eq(44), r_i is the remainder of n_i congruent f.

$$n_i = m_i * f + r_i \quad (44)$$

2) Several 0's can be added to the stream of r_i bits to form a chunk with a length f. The total number of chunks will be equal to $m_i + 1$.
3) The XORing is done by chunks of f bits.
4) The resulting stream of $m_i + 1$ has a deviation to non-randomness that is lower than the incoming stream n_i.

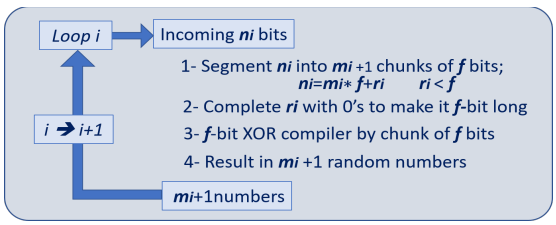


Figure. 19: Generalization of the concept to any data streams.

11. Native ternary random numbers generators

In this section, we are presenting a method to directly generate native random trits from PUF memory arrays, as well as ways to enhance randomness with modulo 3 sum adders. As presented in the first section, the fuzzy cells of the PUFs are used as multiple sources of randomness, and the XOR compilers replace mod3 adders. Ternary computing uses trits, for example (0, 1, 2) or balanced (-, 0, +), instead of the bits (0, 1) used in binary computing [61-67]. Can ternary computing improve cybersecurity and Information Assurance [68-70]? Ternary computing is not a new concept, and is more complex to implement than binary computing. One suggested architecture uses heterogeneous computing elements, and combine binary units to run legacy codes, and native ternary computing units for security [71]:

- Better handling of the natural fuzziness, with lower reliance on error correction codes;
- Can take advantage of ternary hardware, and advances in microelectronics, such as the ternary PUFs described in the first section of this paper [72-79];
- The cryptography based on ternary states has more entropy, and additional levels of freedom to protect both hardware, and software.

For example, let us assume that the length of a data stream is N=128. The number of possible combinations for binary streams is 2^{128}= 3.4 10^{38}, and becomes 3^{128}= 1.2 10^{61} for ternary streams, which is considerably larger. Native random numbers are valuable for cryptographic protocols based on ternary computing. One way to create ternary random numbers is to convert binary random numbers into decimal numbers, then to convert the decimal data stream back into ternary random numbers. Such a method add complexity, and can potentially expose the random numbers to hackers. A direct generation of native ternary random numbers is therefore desirable. The definition of deviation from perfect randomness for ternary TRNG, is similar to the one developed for binary data streams. As presented in the first section, each bit a_i of the perfectly binary random stream $\{a_1, \ldots, a_i, \ldots, a_n\}$ should have precisely the same probability to be either a "1" or a "0". The average deviation from randomness, λ is given by:

$$\mathbf{P}_{(a_i=1)} = \mathbf{P}_{(a_i=0)} = 0.5 \quad (50)$$

$$\lambda = \tfrac{1}{2}\left(|\mathbf{P}_{(a_i=1)} - 0.5| + |\mathbf{P}_{(a_i=0)} - 0|\right)$$

$$= |\mathbf{P}_{(a_i=1)} - 0.5| = |\mathbf{P}_{(a_i=0)} - 0.5| \quad (51)$$

In the case of ternary data streams of trits with "-", "0", and "+" states, the term λ is given by:

$$\lambda = 1/3\left(|\mathbf{P}_{(a_i=-)} - 1/3| + |\mathbf{P}_{(a_i=0)} - 1/3| + |\mathbf{P}_{(a_i=+)} - 1/3|\right) \quad (52)$$

$$0 = \mathbf{P}_{(a_i=-)} + \mathbf{P}_{(a_i=0)} + \mathbf{P}_{(a_i=+)} \quad (53)$$

12. Description of the method

12.1. Segmentation of the fuzzy cells of the memory PUFs

The fuzzy cells , the 0-cells, can be segmented into three subgroups, see Fig. 20:

- The cells that have a higher probability to be tested as "-" are called A-cells. They have an average probability $\mathbf{P}_{A=-}$ to be tested as "-" , $\mathbf{P}_{A=0}$ to be tested as "0", and $\mathbf{P}_{A=+}$ to be tested as "+".
- The cells that have a higher probability to be tested as "0" are called B-cells. They have an average probability $\mathbf{P}_{B=-}$ to be tested as "-" , $\mathbf{P}_{B=0}$ to be tested as "0", and $\mathbf{P}_{B=+}$ to be tested as "+".

$\mathbf{P}_{C=-}$ to be tested as "-" , $\mathbf{P}_{C=0}$ to be tested as "0", and $\mathbf{P}_{C=+}$ to be tested as "+".

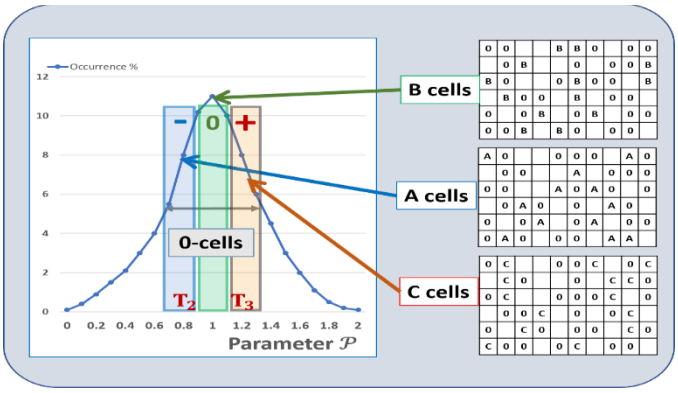


Figure 20: segmentation of the fuzzy cells in trits

The selection of the transition of parameter $\mathcal{P}$ between "-" and "0" , T2, and the transition of parameter $\mathcal{P}$ between "0" and "+", can be such that the total number of A-cells selected within the 0-cells equal the number of B-cells, and the number of C-cells. The deviation from perfect randomness λ of the stream of native ternary random numbers generated from the fuzzy cells A, B, C is given by:

$$\lambda = \frac{|1/9 - \mathbf{P}_{A=-}| + |1/9 - \mathbf{P}_{A=0}| + |1/9 - \mathbf{P}_{A=+}|}{9} + \frac{|1/9 - \mathbf{P}_{B=-}| + |1/9 - \mathbf{P}_{B=0}| + |1/9 - \mathbf{P}_{B=+}|}{9} + \frac{|1/9 - \mathbf{P}_{C=-}| + |1/9 - \mathbf{P}_{C=0}| + |1/9 - \mathbf{P}_{C=+}|}{9} \quad (54)$$

12.2. Enhancement of the randomness with mod3 adders

The algorithm using a mod3 adder, see Fig.21 and 22, is similar to the one presented section A.

1) The number of cells needed to generate a stream of mi trits is $n_i = m_i * f$, they are selected as part of the fuzzy 0-cells of the ternary memory PUF;
2) Parameter $\mathcal{P}$ is measured on all n_i cells;
3) The population of ni cells is segmented into three third based on the value of $\mathcal{P}$. The threshold separating the bottom third and the central third is $\mathbf{T_2}$; the threshold separating the central third and to top third is $\mathbf{T_3}$;
4) With the segmentation in three done step 3, the n_i cells at the bottom third are carrying "-"state, the cells in the middle are "0"s, and the cells at the top third are "+"s;
5) The stream of trits is added by chunks of f trits; Instead of a XOR compiler, the mod 3 adder of chunk of trits enhance randomness. With mod 3 sum adders, two input trits a_i, and a_{i+1} are transformed into $c_i = a_i \oplus a_{i+1}$, with the following truth table:

(a_i =0; a_{i+1} = -), or (a_i = -; a_{i+1} =0), or (a_i=a_{i+1} =+)➔c_i = -

(a_i =+; a_{i+1}= -), or (a_i = -; a_{i+1}= +), or (a_i=a_{i+1} =0)➔c_i = 0

(a_i=0; a_{i+1}= +), or (a_i= +; a_{i+1} =0), or (a_i=a_{i+1} = -)➔c_i = +

6) The resulting m_i trits are more random.

Mod 3 addition increases randomness, the knowledge of c_i is not disclosing the value of a_i and a_{i+1}. c_i = -, **0**, or + can be the result of three possible pairs $\{a_i\ a_{i+1}\}$, with equal probability. If two trits a_i and a_{i+1} are somewhat random, the trit c_i is even more random than either a_i or a_{i+1}. Let us assume that the stream of random trits generated by the 0-cells of the memory PUF array is $\{a_1, a_2, \ldots, a_i, \ldots a_n\}$.

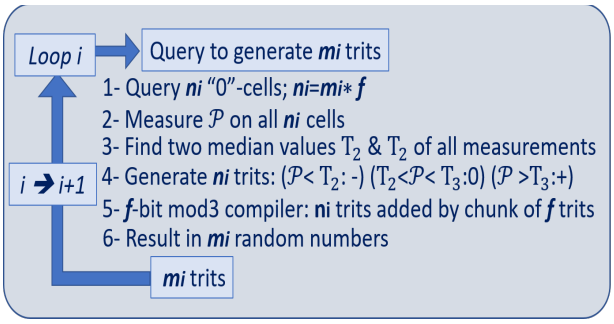


Figure 21: Algorithm to reduce the effect of a parameter drift.

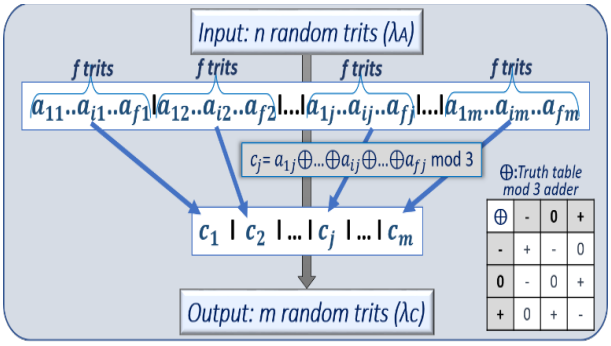


Figure 22: description of the mod 3 adder

As it is shown in Fig.22, this stream is grouped in chunks of f trits $\{a_{1j}, a_{2j}, \ldots, a_{ij}, \ldots, a_{jf}\}$ with $f < n$. For example, 1,280 random bits a grouped in 128 chunks of 10 bits. The resulting stream of random trits obtained with mod 3 sum adders data $\{c_1, c_2, \ldots, c_j, \ldots, c_m\}$ is defined as follow:

$$c_i = a_{1j} \oplus a_{2j} \oplus \ldots \oplus a_{ij} \oplus \ldots \oplus a_{jf} \bmod 3 \qquad (55)$$

Mod 3 sum adders can be implemented at the software level, or in hardware with only a few logic gates. These gates can be inserted in the state machine of the PUF memory to directly feed secure processors, and crypto-processors with streams of randomly generated trits.

13. Modeling of the randomness after mod3 additions

13.1. Model with mod3 addition by chunks of two trits

In this section we are proposing a simplified model that quantifies the level of randomness of mod 3 adders when two adjacent trits are added mod3. Fig.23 shows such a scheme. We are assuming that the 0-cells are distributed into three type of cells (A, B, and C), each of them with a probability of occurrence of 1/3. Statistically the stream of trits $\{a_1, a_2, \ldots, a_i, \ldots a_n\}$ contain trits with equal probability to be "-", "0", or "+", also with a value of 1/3. However A-cells have a higher probability to have "-"s, B-cells have a higher probability to have "0"s, and C-cells have a higher probability to have "+"s.

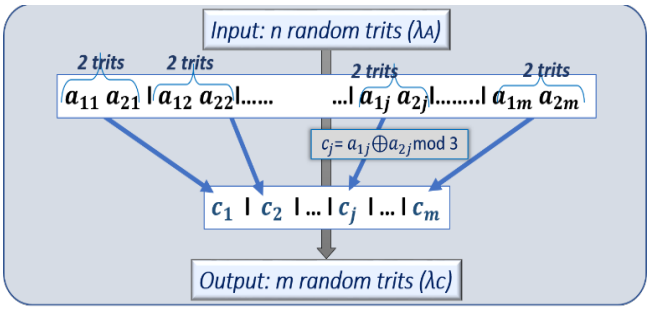


Figure 23: description of the mod 3 adder by chunk of two trits

In Fig.24, we are showing an arbitrary set of probabilities verifying that the probability to have either A, B, or C cells is 1/3, and the probability to have either "-", "0", or "+" states is also 1/3. In this table:

$$P_{A=-} = 1/9 + \Delta_{A-} * \lambda \text{ ; with } \Delta_{A1} = 1.8 \qquad (55)$$

$$P_{B=0} = 1/9 + \Delta_{B0} * \lambda \text{ ; with } \Delta_{A2} = 0.9 \qquad (56)$$

$$P_{C=+} = 1/9 + \Delta_{C+} * \lambda \text{ ; with } \Delta_{C+} = 1.8 \qquad (57)$$

P=1/9+Δ.λ		Δ per type of cell			Prob. to have - ; 0 ; +
		A(-)	B(0)	C(+)	
Prob.Δ when tested	be: -	+1.8	-0.45	-1.35	1/3
	be: 0	-0.45	+0.9	-0.45	1/3
	be: +	-1.35	-0.45	+1.8	1/3
Prob. to have A, B, or C		1/3	1/3	1/3	

Figure 24: example of probabilistic representation

The initial randomness is:

$$\lambda i = (1/9)(|\sum_{i=-}^{i=+} \Delta_{A=i}| + |\sum_{i=-}^{i=+} \Delta_{B=i}| + |\sum_{i=-}^{i=+} \Delta_{C=i}|)$$

$$= 1/9\ (3.6 + 1.8 + 3.6) * \lambda = \lambda \qquad (58)$$

Other tables and more complicated model can replace this arbitrary representation; however, the suggested simplified model describes quite well the experimental observations. When the cells are combined by pairs, and the trits added mod 3, 9 combinations of cells are possible with an equal probability of 1/9: AA, AB, AC, BA, BB, BC, CA, CB, CC. The average probability to have trits with "-", "0", and "+" is 1/3.

Two cells AA have three possible combinations which can result in a trit at "-" :

- Both cells at "+", the probability is: ($P_{A=+} * P_{A=+}$);
- The first cell is at "0", the second at "-" , the probability is: ($P_{A=0} * P_{A=-}$);
- The first cell is at "-", the second at "0", the probability is: ($P_{A=-} * P_{A=0}$).

The resulting probability $P_{AA=-}$ that two cells AA can result in a trit at "-" is given by:

$$P_{AA=-} = (P_{A=+} * P_{A=+}) + (P_{A=0} * P_{A=-}) + (P_{A=-} * P_{A=0})$$

$$= (1/9 - 1.35*\lambda)^2 + 2(1/9 + 1.8*\lambda)(1/9 - 0.45*\lambda) = 1/27 + 0.2*\lambda^2 \quad (59)$$

In Fig.25 is showing the result of the computations of 27 configurations: probability to get "-", "0", or "1" after addition mod 3 for each pair $P_{AA=-}$, $P_{AA=0}$, $P_{AA=+}$, $P_{AB=-}$, $P_{AB=0}$,…

$P=1/27+\Delta.\lambda^2$		Δ per type of pair to form $c = a_1 + a_1$ mod 3									Prob. to have −;0;+
		AA	AB	AC	BA	BB	BC	CA	CB	CC	
Prob.Δ when tested	be: -	+0.20	+2.43	-2.63	+2.43	-0.61	-1.82	-2.63	-1.82	+4.46	1/3
	be:0	-4.66	-0.61	+5.26	-0.61	+1.5	-0.61	+5.26	-0.61	-4.66	1/3
	be:+	+4.46	-1.82	-2.63	-1.82	-0.61	+3.43	-2.63	+2.43	+0.20	1/3
Prob. to have AA, AB,..., or CC		1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	

Figure 25: probability per configuration after mod 3 addition of two trits.

The resulting deviation from randomness $\lambda'f$ after addition is the average deviation of these 27 configurations:

$$\lambda'f = (1/27) \sum |\Delta . \lambda^2| \approx 2.27\, \lambda^2 \quad (60)$$

For example, if the initial deviation from randomness for the incoming stream is $\lambda_i = 2\ 10^{-2}$; the resulting deviation is:

$$\lambda'f = 2.27 * (2 * 10^{-2})^2 = 8.7\ 10^{-4} \quad (61)$$

After addition, the 9 possible configurations shown in Fig.25 can be then combined into 3 types of cells A', B', and C. For example, the cells that are mainly "-", the A'-cells consist of the mod3 additions of CC, AB, and BA pairs. In this case, the average deviation from randomness of A'-cells when they are containing a trit "-"is:

$$(\lambda'f_{A'=-}) = 1/3((\lambda'f_{CC=-}) + (\lambda'f_{AB=-}) + (\lambda'f_{BA=-}))$$

$$= 1/3\ (2.43+2.43+4.46) * \lambda^2 = 3.10 * \lambda^2 \quad (62)$$

13.2. Extension of the model with chunks of four trits

The method presented section II 2.2 can be extended to the addition mod3 of 4 sequential trits to generate trits of higher randomness. Rather than starting with the three types of cells A, B, and C having a deviation from randomness λ, the same computation is done with the cells A', B', and C' having a deviation from randomness equal to $2.27 * \lambda^2$.

The resulting deviation from randomness $\lambda"f$ of the new stream of trits and mod 3 addition of chunks of four bits is:

$$\lambda"f = 2.27\ (\lambda'f)^2 = 2.27\ (2.27)^2\ \lambda^4 \approx 11.3\ \lambda^4 \quad (63)$$

If the initial deviation is $\lambda = 2\ 10^{-2}$; the resulting deviation is:

$$\lambda"f = 11.3 * (2 * 10^{-2})^4 = 1.81\ 10^{-6} \quad (64)$$

By extension, after addition of 8 sequential trits, the deviation from randomness $\lambda""$ will be:

$$\lambda""f = 2.27\ (\lambda"f)^2 \approx 290\ \lambda^8 \quad (65)$$

If the initial deviation is $\lambda = 2\ 10^{-2}$; the resulting deviation is:

$$\lambda""f = 290 * (2 * 10^{-2})^8 = 7.4\ 10^{-12} \quad (66)$$

14. Experimental analysis with mod3 adders

The analysis is based on the data presented in the first section related to the measurement of the Vset of ReRAM devices. We are again considering the same five cases to sort out the fuzzy "0-cells":

- Case-1: Only **2%** of the cells are 0-cells. For these cells parameter $\mathcal{P}$ is very close to the transition of 2.1 Volt. $\lambda_i = 2\ 10^{-2}$;
- Case-2: **4%** of the cells are 0-cells. $\lambda_i = 4\ 10^{-2}$;
- Case-3: **7%** of the cells are 0-cells. $\lambda_i = 6\ 10^{-2}$;
- Case-4: **11%** of the cells are 0-cells. $\lambda_i = 1\ 10^{-1}$;

This range of options allows the quantification of the effectiveness of the addition mod 3 to scramble the trits, as a function of the initial randomness coming from the ternary PUF memory. Case-1 is the one with the higher initial randomness, while Case-4 is the lowest one.

The probabilistic model developed above, is the base of the analysis of the four experimental cases. The results of the computations are shown in Fig.26 and Fig. 27.

The impact of the addition modulo 3 addition on the level of randomness on the resulting data streams of trits is increasing when the number of cells f involved in the addition increases from $f = 2$ to $f = 8$.

Case	% of X-cells	$\lambda i = \lambda$	$\lambda'f \approx 2.27\ \lambda^2$ 2 cells	$\lambda"f \approx 11.3\ \lambda^4$ 4 cells	$\lambda""f \approx 290\ \lambda^8$ 8 cells
1	2%	$2\ 10^{-2}$	$8.7\ 10^{-4}$	$1.81\ 10^{-6}$	$7.4\ 10^{-12}$
2	4%	$4\ 10^{-2}$	$3.6\ 10^{-3}$	$2.9\ 10^{-5}$	$1.9\ 10^{-9}$
3	7%	$6\ 10^{-2}$	$8\ 10^{-3}$	$1.46\ 10^{-4}$	$4.87\ 10^{-8}$
4	11%	$1\ 10^{-1}$	$2.3\ 10^{-2}$	$1.1\ 10^{-3}$	$2.9\ 10^{-6}$

Figure 26: modeling of the effect of mod 3 addition as a function of the experimental cases.

In all cases the addition mod 3 when applied to a data stream of trits generated by a PUF is very efficient to enhance randomness. Case 1, the one with the highest level of initial randomness, benefit the most from mod 3 sum adders.

It is interesting to notice that the XOR compiler, and the mod 3 addition have similar effects in improving randomness. The model can be used as a predictive tool to anticipate the level of randomness of streams of trits. For example, as shown in Fig.27, the number cells necessary to get $\lambda < 5\ 10^{-6}$ for case 1 is 4, it is 5 for case 2, it is 6 for case 3, and it is higher than 8 for case 4.

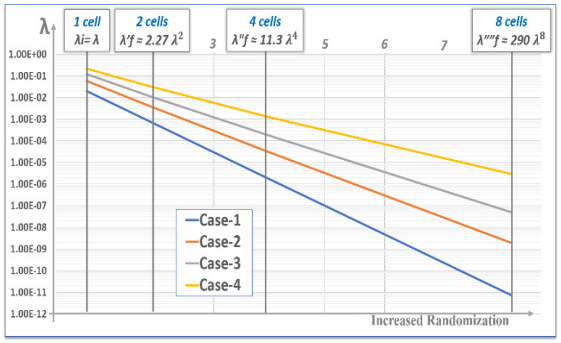


Figure 27: deviation from perfect randomness plotted as a function of the size of the chunk of trits involved in mod3 addition for the four cases.

Discussion and conclusion

The use of ternary PUFs to design TRNG, combined with a XOR compiler, or a modulo 3 addition has the following benefits:

- The cells with fuzzy behavior of a ternary PUF, the 0-cells, can provide multiple sources of independent randomness. A memory arrays in the megabyte range can have a large quantity of such cells;
- The randomness of the binary data streams extracted from the ternary 0-cells are enhanced by a XOR compiler. Based on a normal distribution of parameter $\mathcal{P}$, the proposed statistical model can quantify the deviation from pure randomness of the TRNGs. It is possible to calculate f, the length of the XOR, to reach a desired level of non-randomness λ_f, as a function of the level of non-randomness λ_i of the incoming data stream extracted from the physical element.
- The randomness of the ternary data streams extracted from the 0-cells can be enhanced by a modulo 3 adder. It is the possible to directly generate a stream of random trits without having to convert binary data streams into ternary data streams;
- It is possible to anticipate, with the suggested probabilistic model, the minimum size of chunks of data f that need to be processed to reach the level of randomness λ_f. This is the case for the XOR compiler, and the mod 3 adder;
- The proposed methodology minimizes sensitivity to parameter drifts such as temperature, aging, or biasing conditions. It is anticipated that the drifts should not materially degrade the quality of the TRNGs.
- The hardware implementation of both the XOR compiler, and the mod 3 adder can use known commercial CMOS circuitry.
- The method can reach NIST expectations in term of deviation from pure randomness of the TRNG, even if the randomness created by the PUF is weak.

The experimental section of this work, which is based on the measurements of the Vset of ReRAM cells, produced a distribution that is able to show enough randomness to generate random numbers. We noticed that the XOR data compiler is not effective when the initial data stream is not random. The model developed assumes that the initial random numbers generated from the 0-cells are symmetrically distributed between A-cells and B-cells.

Other statistical distributions beside the normal one are under consideration in our research effort. We are not anticipating that these improved statistical models will significantly change the outcome when only cells close to the median distribution are selected. This is not the case for wider distribution of the 0-cells away from the median.

Future work: The objective of this work was to develop TRNG for cryptographic protocols that can be embedded in the Internet of Things (IoT). The implementation of affordable sources of randomness to secure IoTs can benefit from the ease of use of ternary PUFs, which are tamper resistant. TRNGs are essentials elements to encryption protocols involving PUF CRPs, and other cryptographic keys. We are studying the design of a prototype that incorporates the proposed TRNG scheme with various ReRAM arrays. The prototype is intended to automatically extract large quantities of PUF CRPs and random streams of bits, and trits. We intend to use the prototype to further validate our statistical models, and to leverage the tools developed by NIST that are available online to quantify the entropy, and the level of randomness of the TRNGs. The prototype should have the built-in flexibility to allow us to analyze multiple types of memory arrays, with different methods of fabrication.

We are interested in optimizing the randomness of the TRNG while reducing CRP error rates of the PUFs, and developing cryptographic protocols that leverage the combined capabilities. The method described in this paper can be used to the swarm dynamics generating true random noises [8], and other similar applications requiring TRNG. To accelerate the process to generate fresh random numbers on demand, the 0-cells can be tested in advance [80], and the data can be stored in the memory. The read time of a ReRAM is typically 10ns/bit, so we believe that the generation of the TRNG has the potential to be done at a rate of 100Mbit/s. The method presented in this paper can also be extended to n-value logic, for example quaternary logic (4 bits), pentagonal logic (5 bits), or hexagonal logic (6bits). In such cases the 0-cells are divided in n different type of cells, and the addition of chunk of n-bits is done modulo n.

Aknowledgment

Dr. Marius Orlowski from Virginia Tech produced, and characterized the metal oxide ReRAM samples [49]. These measurements were used to prepare the experimental analysis. Dr. Derek Sonderegger from the Department of Mathematics and Statistic of Northern Arizona University provided input for the development of the statistical model. The author is thanking both Dr. Donald Telesca from the US Air Force Research Labs for his support in this work.

References

[1] B. Cambou; A XOR data compiler combined with PUF for TRNG; SAI/IEEE computing conference, July 2017;
[2] C. Paar, and J. Pezl; Understanding Cryptography- A text book for students and practitioners; Spinger editions, 2011;
[3] C. P. Pfleeger, et al; Security in Computing; Fifth edition; Prentice Hall editions, 2015;
[4] R. Soorat; Hardware random number generation for cryptography; https://arxiv.org/pdf/1510.01234 , 2015;
[5] D. Glosemeyer, and B Knapp; Random Number Generation; Wolfram Mathematical Tutorial collection, 2008;
[6] M. Stipevic, et al; true random number generator; Open problems in Mathematics and computational science; pp.275-315, Springer, 2014;
[7] H. Katzgraber; Random Numbers in scientific computing: an introduction; Int. school comp. science, 2010, Oldenburg, Germany
[8] Y. Shang, and R. Bouffanais; "Influence of the number of topologically interacting neighbors on swarm dynamics"; Scientific Reports DOI: 10.1038/srep04184, 2015.
[9] Berndt Gammel, et al; Jul 2012; Random Generator configured to combine states of memory cells; US patent No 7,979,482 B2;
[10] Daniel E Holcomb, et al; Power up SRAM state as an identifying Fingerprint and Source of True Random Numbers; IEEE Trans. On Computers, 2009 Vol 58, issue No09 Sept;

[11] Sang-Yong Yoon, et al; Aug 2012; Method of operating nonvolatile memory devices storing randomized data generated by copyback operation; US patent No: 8990481 B2;

[12] J. Soto; Statistical testing of random number generators; 1999 – NIST; http://csrc.nist.gov/rng/rng5.html ;

[13] A. Rukhin, and all; A statistical Test Suite For Cryptographic Applications; NIST publication 800-22rev1a, April 2010.

[14] P. L'Ecuyer; Software for uniform random number generation: distinguishing the good from the bad; 2001 Simulation Conference;

[15] A. Rukhin, et all; A statistical test suite for random and PRNG for cryptographic applications; 2010, publication from NIST 800-22 rev 1a;

[16] A. Maiti, et al; PUF and TRNG: a compact and scalable implementation; GLSVLSI'09, May 2009, Boston;

[17] B. Cambou; Sept, 2015; Random numbers generator with ternary memory; US patent application 2017-0046129;

[18] G. Marsaglia; XOR shift RNG; Journal of Statistical software, 2003.

[19] M. C Tzannes, A. Friedmann; Nov 2001; Randomization using an XOR scrambler in multicarrier communications; US patent No 9191939 B2;

[20] R. Davies; XOR and hardware random number generator; http://www.robertnz.net/pdf/XOR2.pdf; February 2002;

[21] B. Cambou; Data compiler for True Random Number Generation and Related Methods; NAU disclosure D2017-03, Aug 2016;

[22] Y. Gao, and all; Emerging Physical Unclonable Functions with nanotechnologies; IEEE, DOI: 10.1109/ACCESS.2015.2503432;

[23] N. Beckmann, et al; Hardware-based public-key cryptography with public physically unclonable functions; in Information Hiding, New York, NY, USA: Springer-Verlag, 2009, pp. 206–220.

[24] Guajardo, J, et al; PUFs and PublicKey Crypto for FPGA IP Protection; Field Programmable Logic and Applications, 2007.

[25] David. Naccache and Patrice. Frémanteau; Aug. 1992; Unforgeable identification device, identification device reader and method of identification; Patent US5434917.

[26] Guajardo, J., et al; FPGA Intrinsic PUFs and Their Use for IP Protection; CHES, 2007;

[27] R. Pappu, B. Recht, J. Taylor, and N. Gershenfield; 20 Sept 2002; Physical one-way functions; Science. Vol 297 No5589 pp2026-2030.

[28] Yier Jin; Introduction to hardware security, Electronics 2015, 4, 763-784; doi:10.3390/electronics4040763.

[29] R. Maes; Physically Unclonable functions: constructions, properties, and applications; Doctoral thesis- Catholic University of Leuven, 2012;

[30] Herder, C., et al; "PUFs and Applications; A Tutorial." Proceedings of the IEEE 102, no. 8 (2014): 1126-1141;

[31] B. Gassend, et al; Silicon PUFs; CCS' 2002;

[32] B. Gassend; Physical random functions; M.S. thesis, Dept. Electr. Eng. Comput. Sci., MA, USA, Massachusetts Inst. Tech.., Cambridge, 2003;

[33] S. Katzenbeisser, et al; PUFs: myths, fact or busted? A security evaluation of PUFs cast in silicon; CHES 2012;

[34] Christian Krutzik; Jan 2015; Solid state drive Physical Unclonable Function erase verification device and method; US Patent Application publication US 2015/0007337 A1

[35] H. Kang, Y. Hori, T. Katashita, M. Hagiwara, and K. Iwamura; Cryptographic Key Generation from PUF Data Using Efficient Fuzzy Extractors; in Proc. ICACT, 2014, pp.23–26.

[36] R. Maes, P. Tuyls and I. Verbauwhede, "A Soft Decision Helper Data Algorithm for SRAM PUFs," in 2009 IEEE International Symposium on Information Theory, 2009.

[37] Daniel E. Holcomb, Wayne P. Burleson, Kevin Fu; Nov 2008; Power-up SRAM state as an Identifying Fingerprint and Source of TRN; IEEE Trans. on Comp., vol 57, No 11.

[38] D. E. Holcomb, and all; Power-up SRAM state as an Identifying Fingerprint and TRN; IEEE Trans. Comp. Nov 2008;

[39] Pravin Prabhu, Ameen Akel, Laura M. Grupp, Wing-Kei S. Yu, G. Edward Suh, Edwin Kan, and Steven Swanson; June 2011; Extracting Device Fingerprints from Flash Memory by Exploiting Physical Variations; 4th int. conference on Trust and trustworthy computing;

[40] V.Zhirnov, et al; Chapter 26: Flash memories; Nanoelectronics and Information Technology; Rainer Waser editor, 2012 Wiley;

[41] T. A. Christensen, and all; PUF utilizing EDRAM memory cell capacitance variation; Patent: US 8,300,450B2; Oct, 2012;

[42] T. A. Christensen, J. E Sheets II; 2012; Implementing PUF utilizing EDRAM Memory Cell Capacitance Variation; US Patent 8,300,450 B2.

[43] K.K. Chang et al; Understanding Reduced-Voltage Operation in Modern DRAM Chips: Characterization, Analysis, and Mechanisms; Cornell Technical Library: 1705.102992, May 2017;

[44] U. Schroder, et al; Capacitor-based Random-Access Memories; Nanoelectronics and information technology; Wiley-vch, R. Waser editor, pp 635-654, 2012;

[45] Xiaochun Zhu, Steven Millendorf, Xu Guo, David M. Jacobson, Kangho Lee, Seung H. Kang, Matthew M. Nowak, Daha Fazla; March 2015; PUFs based on resistivity of MRAM magnetic tunnel junctions; Patents. US 2015/0071432 A1.

[46] Elena I. Vatajelu, Giorgio Di Natale, Mario Barbareschi, Lionel Torres, Marco Indaco, and Paolo Prinetto; July 2015; STT-MRAM-Based PUF Architecture exploiting Magnetic Tunnel Junction Fabrication-Induced Variability; ACM transactions.

[47] A. Chen; Comprehensive Assessment of RRAM-based PUF for Hardware Security Applications; IEDM IEEE; 2015;

[48] B.Cambou; Enhancing Secure Elements- Technology and Architecture; Springer Int. Publishing Foundations of Hardware IP Protection, 2017;

[49] B. Cambou, and M. Orlowski; PUF designed with ReRAM and ternary states; CISR 2016, April 2016, Oak ridge;

[50] D. Yamamoto, K. Sakiyama, K. Ohto, and M. Itoh; Uniqueness Enhancement of PUF Responses Based on the Locations of Random Outputting RS Latches; CHES 2011;

[51] B. Cambou; PUF generating systems and related methods; US patent disclosure No: 62/204912; Aug 2015;

[52] B. Cambou, and F. Afghah; PUF with Multi-states and Machine Learning; CryptArchi 2016

[53] Gilbert, Nad, et al. "A 0.6 V 8 pJ/write Non-Volatile CBRAM Macro Embedded in a Body Sensor Node for Ultra Low Energy Applications." VLSI Circuits (VLSIC), 2013 Symposium on. IEEE, 2013;

[54] M. N. Kozicki, M. Park, and M. Mitkova, "Nanoscale memory elements based on solid-state electrolytes," IEEE Trans. Nanotechnol, vol. 4, pp. 331-338, May 2005;

[55] M. N. Kozicki and M. Mitkova, "Mass transport in chalcogenide electrolyte films – materials and applications," J. of Non-Crystalline Solids, vol. 352, pp. 567-577, March 2006;

[56] Valov, R. Waser, J. R. Jameson and M. N. Kozicki, "Electrochemical metallization memories - Fundamentals, applications, prospects," Nanotechnology, vol. 22, p. 254003, June 2011;

[57] M. N. Kozicki, M. Balakrishnan, C. Gopalan, C. Ratnakumar, and M. Mitkova, "Programmable metallization cell memory based on Ag-Ge-S and Cu-Ge-S solid electrolytes," Proc. NVMTS, p. 8389, 2005;

[58] M. N. Kozicki, C. Gopalan, M. Balakrishnan, M. Park and M. Mitkova, "Nonvolatile memory based on solid electrolytes," in Proc. IEEE Non-Volatile Memory Technol. Symp., 2004;

[59] Gargi Ghosh and Marius Orlowski; 2015; Write and Erase Threshold Voltage Interdependence in Resistive Switching Memory Cells; IEEE trans. on Electron Devices, 62(9), pp. 2850-2857.

[60] P. R.Mickel, A. J. Lohn, B. J. Choi, J. J. Yang, M. X. Zhang, M. J. Marinella, C. D. James, and R. S. Williams, "A physical model of switching dynamics in tantalum oxide memristive devices," Appl. Phys. Lett., vol. 102, p. 223502, 2013;

[61] M. Glusker, D.M. Hogan, and P. Vass; The ternary calculating machine of Thomas Fowler; IEEE Annals of the History of Computing, 2005;

[62] N.P. Brousentov, S.P. Maslov, J. Ramil Alvarez, and E.A. Zhogolev; Development of Ternary computers at Moscow State University; Russian Virtual Computer Museum; 1997-2017;

[63] G. Frieder; Ternary Computers, part 1: motivation for ternary computers; Micro 5 Conf. record of the 5th annual workshop on Microprogramming; 1972;

[64] S. Ahmad, M. Alam; Balanced Ternary Logic For improving Computing; IJCSIT, 2014;

[65] I. Profeanu; A ternary Arithmetic and Logic; WCE, June 2010;

[66] M.G. Nektar, D.M. Hogan, P. Vass; The ternary calculating machine of Thomas Fowler; IEEE Annals of the History of Computing, Aug 2005;

[67] E.W. Dijkstra; Notes on structured programming; EWD 249 Technical University, Eindhoven, Netherlands, 1969;

[68] B. Cambou, P. Flikkema, J. Palmer, D. Telesca, C. Philabaum; Can Ternary Computing Improve Information Assurance?; MDPI, Journal cryptography, Feb 2018;

[69] D.M. Miller, and M.A. Thornton; Multiple Valued Logic: Concepts and Representations; Synthesis Lectures on Digital Circuits and Systems; Morgan & Claypool Publishers, 2007;
[70] S. Ahmad, and M. Alam; Balanced Ternary Logic For improving Computing; IJCSIT, 2014;
[71] P.G. Flikkema and B. Cambou; Adapting Processor Architectures for the Periphery of the IoT Nervous System; IEEE 3rd World Forum on Internet of Things (WF-IoT), December 2016;
[72] H. Gundersen; Aspect of balanced ternary arithmetic implemented using CMOS recharged semi-floating gate device; thesis Oslo Univ. 2008;
[73] P.C. Balla, and A. Antoniou; Low Power Dissipation MOS Ternary Logic Family; IEEE J of Solid State Circuits, Oct 1984;
[74] P.C. Balla, A. Antoniou; Low Power Dissipation MOS Ternary Logic Family; IEEE J of Solid State Circuits, Oct 1984;
[75] X.W. Wu; CMOS Ternary Logic Circuits; IEE Proceedings, Feb 1990
[76] P. Nagaraju, et al; Ternary Logic Gates and Ternary SRAM Implementation in VLSI; IJSR, Nov 2014;
[77] N.P. Wanjari, S.P. Hajare; VLSI Design and Implementation of Ternary Logic Gates and Ternary SRAM Cell; IJECSE, April 2013;
[78] A. Srivastava, and K. Venkatapathy; Design and Implementation of a Low Power Ternary Full Adder; VLSI design, vol 4, No.1, 1996;
[79] N.P. Wanjari, and S.P. Hajare; VLSI Design and Implementation of Ternary Logic Gates and Ternary SRAM Cell; IJECSE, April 2013;
[80] Anuj Gupta, May 2005, Implementing Generic BIST for testing Kilo-Bit Memories; Master Thesis No-6030402 Deemed University Patiala India

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 16-22 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Improvement of Transmission Characteristics in Multilayer Dual Band Filter

Ryosuke Nakano*, Yuta Takeuchi, Mikio Tsuji, Hiroyuki Deguchi

Department of Electronics, Doshisha University 1-3, Miyakodani, Tatara, Kyotanabe-shi, Kyoto, 610-0321 Japan

ARTICLE INFO

Article history:
Received: 02 July, 2018
Accepted: 20 August, 2018
Online: 08 September, 2018

Keywords:
Genetic algorithms
Band-pass filter
Ultra wide band
Dual-band

ABSTRACT

This paper first simplifies the GA-optimized filter structure developed by us. Then the filter structure is modified by introducing slot and conductor stubs, based on electric fields and magnetic currents at resonant and transmission zero points. As a result, return losses in two passband regions are suppressed more than 28 dB, and also insertion loss in stopband dividing two passbands is suppressed more than 20 dB. Finally, two kinds of the proposed filters are fabricated, and effectiveness of them is confirmed from comparison between the measured and the numerical results.

1. Introduction

This paper is an extension of work originally presented in 2017 Asia Pacific Microwave Conference, (APMC2017) [1]. Based on our work in [2], the goal is to develop high-performance UWB bandpass filter.

In recent years, wireless communication systems such as wireless local area network (LAN) and mobile terminals have spread rapidly. High-frequency filters are used to extract a signal of a desired frequency band for the purpose of prevention of unnecessary radiation and interference. Increasing in complexity and volume of wireless communication, a filter having a large attenuation amount close to pass band and an improved insertion loss is indispensable. Conventional resonant elements that have been developed to achieve high performance such as improved multi-band sharing and attenuation characteristics outside a pass band required for the wireless communication device are not sufficient. Therefore, development of a new resonant element desired. Under such circumstances, we have proposed the ultra-wideband (UWB) filter, of which conductor patches and slot shapes were optimized under the genetic algorithm (GA) technique [2]. As for UWB filters, many filter structures have been developed [3-17]. They are essentially based on the multi-mode resonators (MMR) which were first developed in [3]. Then, various stub-loaded resonators were often used to achieve good skirt selectivity and compact size [4-7]. To perform further good filter performance, the resonator shapes including a stepped-impedance technique were investigated [8-17]. The planar circuit filter designed by GA shows the desired characteristics according to the evaluation function, but the return loss there was not so high (about 10 dB). In addition to the problem of calculation capacity and computation time, the circuit element shape is complicated and the degree of freedom is small, so that it is difficult to improve the filter characteristics by slightly modifying the GA-optimized filter shape.

In this work, we clarify resonant mechanism and transmission zero mechanism of the UWB filter optimized by the GA. Then, the simplified filter structure based on the their mechanisms is proposed. Furthermore, to improve loss characteristics between two passbands of the dual-band filter, the conductor and slot stubs are newly loaded. Validity of the proposed filter is also verified experimentally.

The rest of this paper is structured/organized as follows: Section 2 discusses resonant mechanism of the UWB filter; Section 3 presents improvement of filter transmission characteristics; Section 4 verifies effectiveness of the proposed filter; and finally, in Section 5, we summarize the conclusions derived from this work and indicate possible future works.

2. Simplified filter generated by GA

New planar circuit filters have been developed by GA evaluated characteristics of both passband and stopband [18-19]. In this section, we briefly describe the structure of the multilayer planar circuit filter and the outline of the optimized design by the GA of the planar circuit filter [18-19]. After the circuit shape and filter characteristics are described, we simplify the filter based on resonant principle.

*Corresponding Author: Ryosuke Nakano, ctwb0339@mail4.doshisha.ac.jp
www.astesj.com
https://dx.doi.org/10.25046/aj030503

2.1. Structure of the GA-optimized UWB filter

Figure 1 shows the UWB-filter structure developed by us. This filter optimized by GA consists of slot elements on the ground plane and patch elements on the conductor-backed dielectric substrate ($\varepsilon_r = 2.2$, thickness $t_2 = 0.51$ [mm]). Both elements have arbitrary shapes and the filter is covered by the shielding box ($a = b = 15.2$ [mm], $c = 20.0$ [mm]). Input and output of the filter are conected through a microstrip line. A lot of resonances are mainly caused by slot elements.

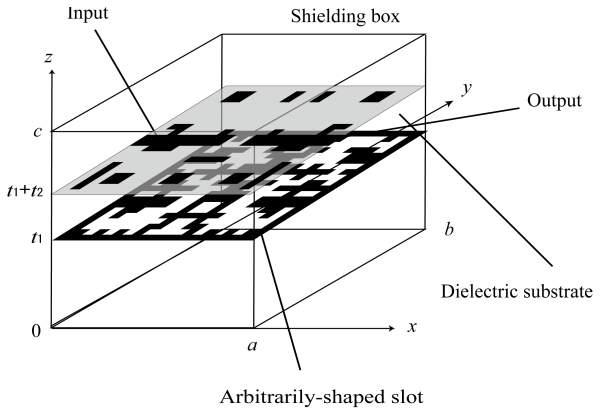


Figure 1. Structure of the GA-optimized UWB filter.

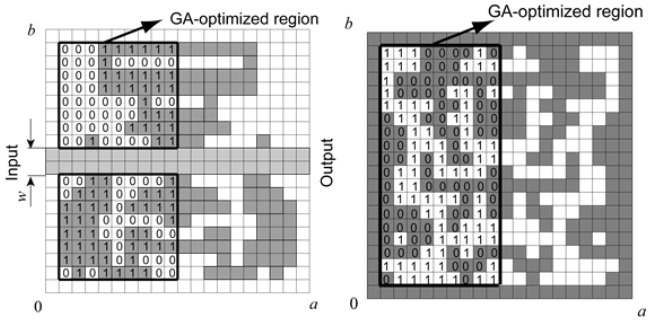
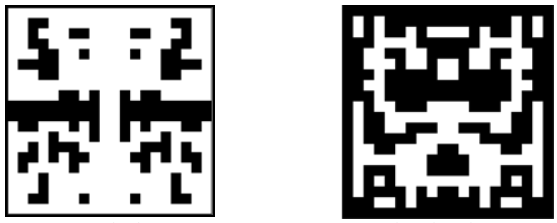


Figure 2. An example of an arbitrarily-shaped element constructed by uniform mesh.

2.2. Shape of the GA-optimized UWB filter

As shown in Fig. 2, the planar circuit shape is represented by a set of conductor patch cells and handled by binary codes of "0" and "1" in the optimized design method by GA [3]. The analysis is performed by the moment method. In the design, the symmetrical structure on the yz plane of $x = a/2$ is assumed and the shape of the input/output line is fixed. The parameters for optimization by GA are set as follows. The xy planes at $z = t_1$ and $t_1 + t_2$ are divided into 20×20 cells, and the size of one cell is 0.76 mm. The evaluation frequency bands are 3.4–4.8 GHz, and 7.25–10.25 GHz in the passband, while 1.0–3.0 GHz, 5.0–7.0 GHz, and 10.35–14.0 GHz in the stopband, and the target of the return loss in the passband and the insertion loss in the stopband are set as 10dB. The mutation rate is 0.02 and the crossover rate is 0.8. The generation number is 400, the population size of one generation is 50, and the elitist strategy is applied.

Figure 3 (a) and (b) show the optimized patch and slot shapes, respectively and Fig. 4 shows its transmission characteristics. It is found that the passbands in the low and high frequency ranges are constituted by five resonances. Since the circuit shape obtained from GA is very complicated, it is very difficult to modify its shape for improving the filter characteristic. Therefore, the resonant principle is clarified from the electric-field distribution, and the shape is simplified based on this.



(a) Patch shape (b) Slot shape

Figure 3. Shape of the GA-optimized UWB filter.

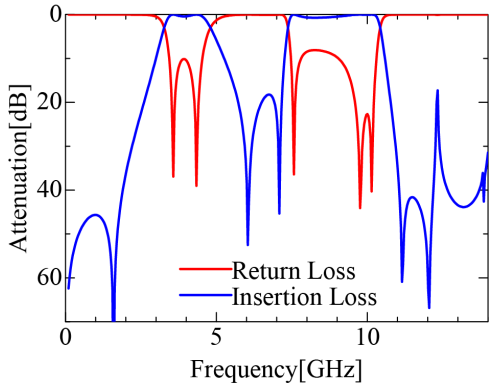
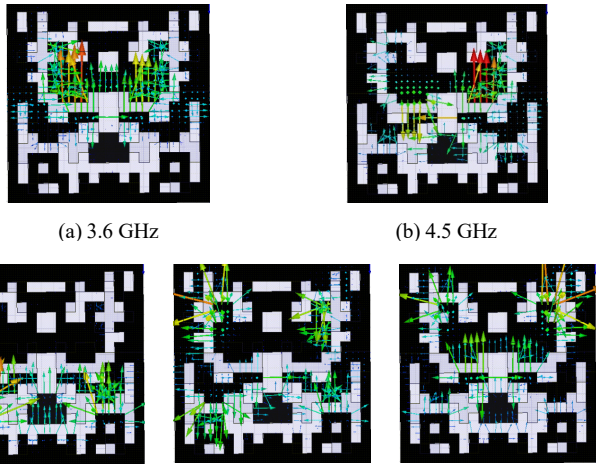


Figure 4. Frequency characteristics of the UWB filter optimized by GA.



(a) 3.6 GHz (b) 4.5 GHz

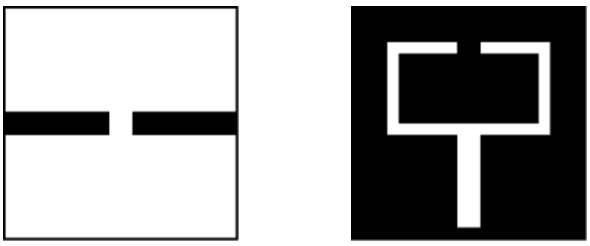
(c) 7.6 GHz (d) 9.6 GHz (e) 10.1 GHz

Figure 5. Electric-field distribution at each resonance.

2.3. Resonant mechanism

Resonances are mainly caused on the slot face, so we observe the electric-field distributions on the slots. Figure 5 shows the electric field on the slot at each resonant point. Resonance at 3.6 GHz shown in Fig. 5 (a), is the half-wavelength even resonant mode of the upper-part slot. Resonance at 4.5 GHz in Fig. 5 (b) is the half-wavelength odd resonant mode of the two slots arranged symmetrically for the center line. Resonance at 7.6 GH in Fig. 5 (c) is the half-wavelength even resonant mode of the lower-part slot. Resonance at 9.6 GHz in Fig. 5 (d) is the odd resonant mode that is the next higher-order one of the resonance at 4.5 GHz. Resonance at 10.1 GHz in Fig. 5 (e) is the 3/2-wavelength even

resonant mode that is the higher-order one of the resonance at 3.6 GHz. From the above-mentioned guess of the resonant mechanism, we first calculate the frequency characteristics for the simplified structure shown in Fig. 6 (a) and (b). Its frequency characteristic is shown in Fig. 7, where four



(a) Patch shape (b) Slot shape

Figure 6. Basic structure of the upper-part slot.

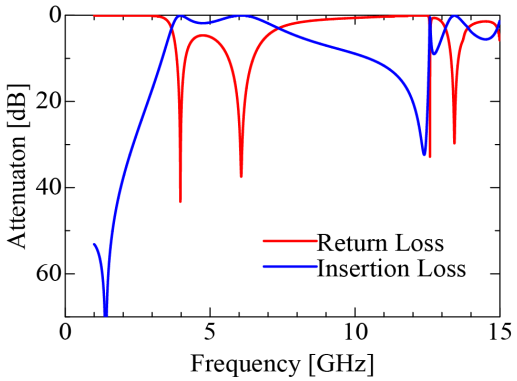
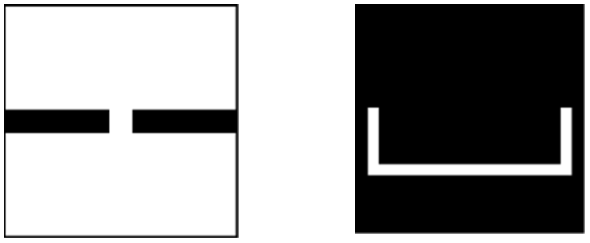


Figure 7. Frequency characteristic of the simplified slot structure shown in Fig.6.



(a) Patch shape (b) Slot shape

Figure 8. Basic structure of the lower-part slot.

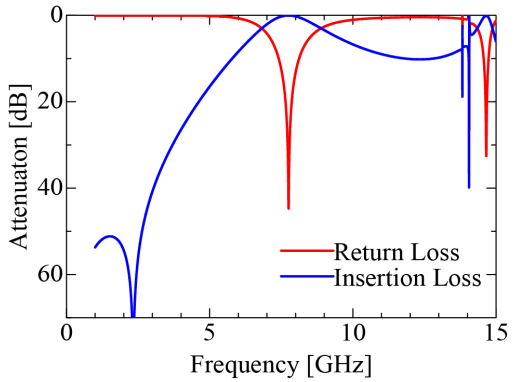
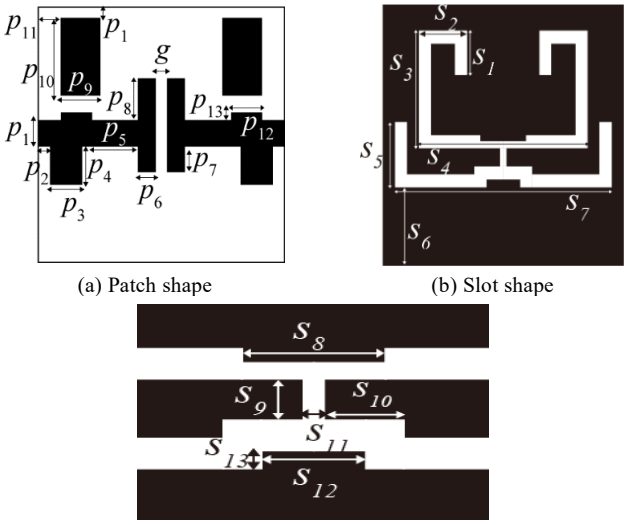


Figure 9. Frequency characteristic of the simplified slot structure shown in Fig.8.

resonant points are produced corresponding to the first, the second , the forth and the fifth resonant points in Fig. 4. Figures 8 (a) and (b) also show another simplified structure and its frequency characteristic is given in Fig. 9, where one resonant point is produced corresponding to the third resonance in Fig. 4. Then we combine these two basic simplified structures and adjust the filter characteristics to the UWB filter by adding slot stubs and microstrip patches and also by modifying the slot junction part. As a result, Fig. 10 (a) and (b) are obtained as a simplified filter structure and the enlarged part of the slot junction is also shown in Fig.10 (c) The dimensions are listed in Table 1. The slot width is 0.76 mm. Figure 11 shows the frequency characteristics of the simplified UWB filter. Five resonances are reproduced near the original resonant frequencies shown in Fig. 4. Figure 12 shows the magnetic currents at the resonant point to confirm the resonant mechanism. It is clear from this figure that the simplified filter structure works by the same resonant mechanism with that of the GA optimized filter. However, the insertion loss in the stopband is not so high less than 20 dB. So we improve it by considering mechanism of the transmission zeros in the next session.

Table 1. Dimensions of the patch and the slot parameters (unit: mm).

$p_1 = 1.52$	$p_2 = 0.76$	$p_3 = 1.78$	$p_4 = 2.24$	$p_5 = 3.61$
$p_6 = 1.04$	$p_7 = 1.52$	$p_8 = 2.78$	$p_9 = 2.38$	$p_{10} = 4.90$
$p_{11} = 1.42$	$p_{12} = 0.74$	$p_{13} = 2.58$	g = 0.74	$s_1 = 2.58$
$s_1 = 3.04$	$s_2 = 6.84$	$s_3 = 10.64$	$s_4 = 3.80$	$s_5 = 4.56$
$s_6 = 13.68$	$s_7 = 5.80$	$s_8 = 1.06$	$s_9 = 1.47$	$s_{10} = 1.4$
$s_{11} = 3.06$	$s_{12} = 0.46$	The slot width is 0.76 mm		



(a) Patch shape (b) Slot shape

(c) Slot shape of the slot-junction point

Figure 10. Shape of the simplified UWB filter.

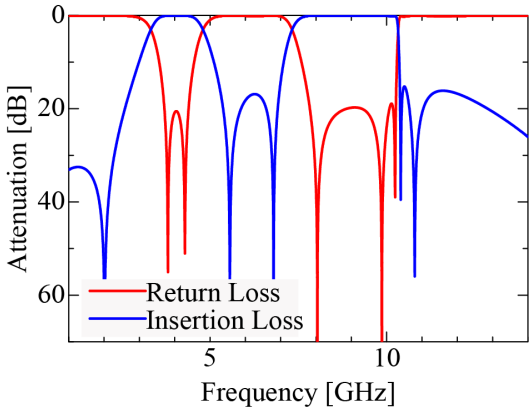
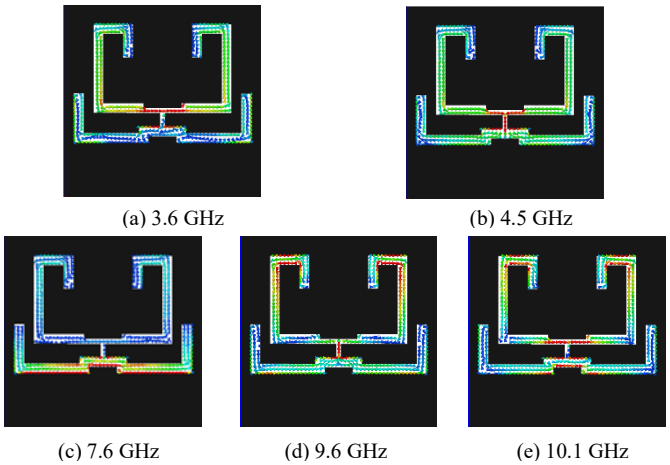


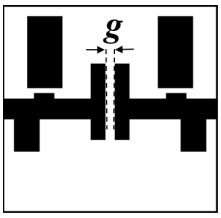
Figure 11. Frequency characteristics of the simplified filter.



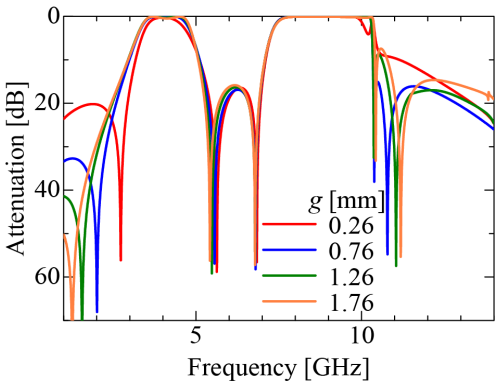
(a) 3.6 GHz (b) 4.5 GHz

(c) 7.6 GHz (d) 9.6 GHz (e) 10.1 GHz

Figure 12. Magnetic-current distribution at each resonance.



(a) patch shape



(b) Insertion Loss

Figure 13. Frequency for various gap g.

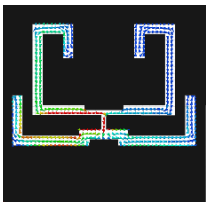


Figure 14. Magnetic-current distribution at 5.5 GHz.

3. Improvement of stopband characteristics

We can see five transmission zeros within 12 GHz in Fig. 4. Among these transmission zeros, the first and second zeros can control by slight modifications of the filter shape, which have little influence for the passband characteristics. We first take notice of the transmission zero at 2.0 GHz. This transmission zero may be caused by the cancellation due to the direct coupling between the input and the output microstrip line through the gap g shown in Fig. 13 (a). Figure 13 (b) shows the transmission characteristics for varying the value of the gap g. As expected, only the first transmission zero greatly moves, depending on g. It is clear that the gap g is a useful parameter for controlling the frequency point

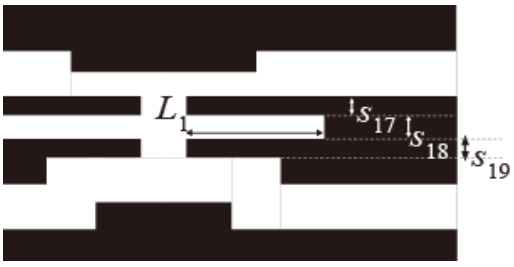
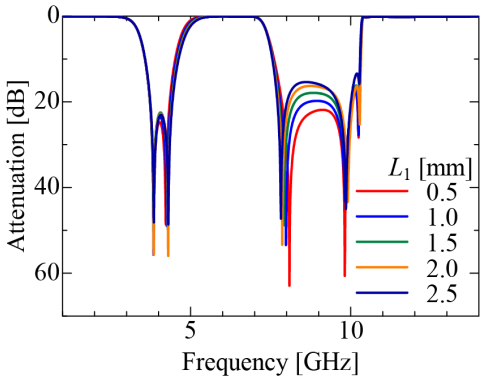
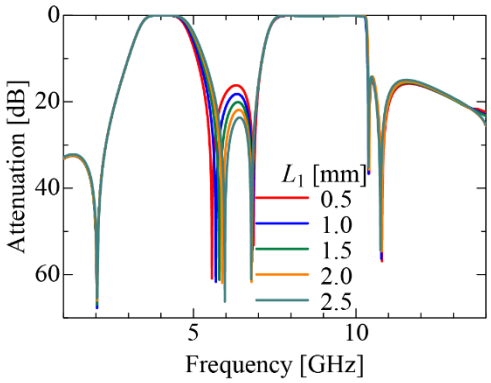


Figure 15. Parameters of the slot stub.



(a) Return loss

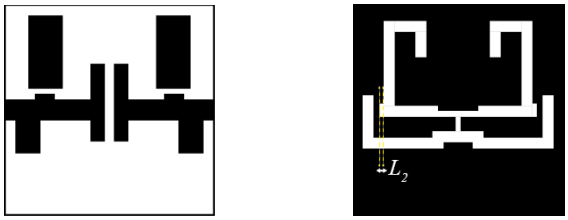


(b) Insertion loss

Figure 16. Frequency characteristics for various slot stub length L_1.

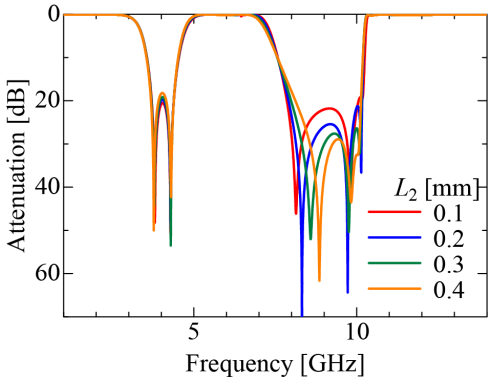
of the first transmission zero. Next, we notice of the second transmission zero of 5.5 GHz. Figure 14 shows the magnetic-current distribution at 5.5 GHz. We can see from this figure that the magnetic current strongly flows in the center slot connecting between the lower-part and higher-part slots. Therefore, if the slot stub is loaded into the center slot, the transmission zero at 5.5 GHz can move independently.

Figure 15 shows the slot stub of the length L_1 and the other parameters are $s_{17} = 0.33$ [mm], $s_{18} = 0.30$ [mm], $s_{19} = 0.33$ [mm]. Figure 16 shows the frequency characteristics for varying L_1. Although only the second zero point shifts to the higher frequency region with increasing L_1 and the stopband characteristic is improved as expected, the third resonant point also moves to the higher region, and the passband characteristic slightly deteriorates. The third resonance at 7.6 GHz is generated by the bottom slot as mentioned in Sec. 2, but if the length of the resonator is made longer, the second and the forth resonances also a little bit move, so such a deformation of the bottom slot for characteristic improvement is difficult. Therefore, we install a small projection of the length L_2 in the upper slot resonator shown in Fig. 17. The projection is coupled with the lower slot resonator so that only the

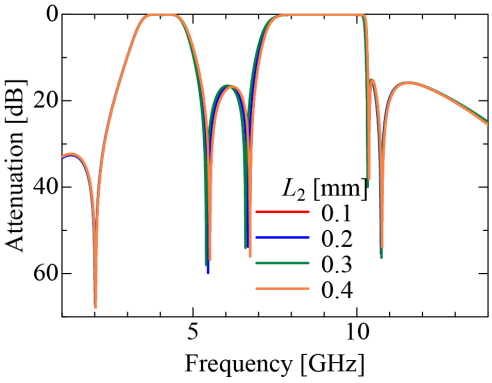


(a) Patch shape (b) Slot shape

Figure 17. Shape of the filter with an additional projection of the length L_2

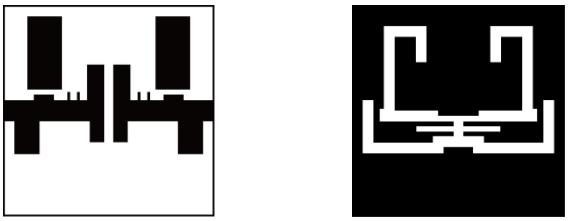


(a) Return loss



(b) Insertion loss

Figure 18. Filter characteristics for various projection length L_2.



(a) Patch shape (b) Slot shape

Figure 19. Proposed filter structure loading two conductor stubs.

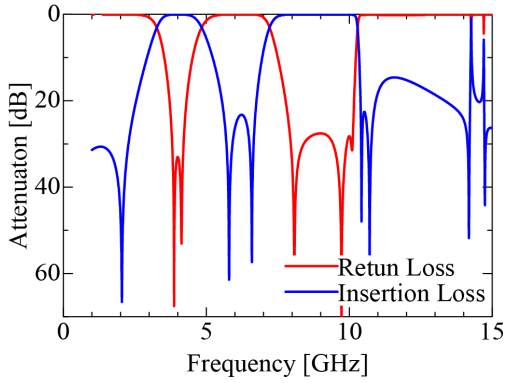


Figure 20. Filter characteristics of the proposed filter.

third resonance can move. Figure 18 shows the transmission characteristics for varying value of L_2. Clearly only the third resonance moves and other resonances, and the transmission zero points don't move. Furthermore, two small conductor stubs are loaded on the input/output line in order to improve the transmission characteristics at around 10 GHz as shown in Fig. 19. Figure 20 shows the transmission characteristics of the filter proposed under the above-mentioned investigation and, the return losses in the two passbands are suppressed to 28 dB, and the insertion loss in the stopband is suppressed to 20 dB or more. Most of UWB filters do not reach to these-losses levels in three pass and stop bands.

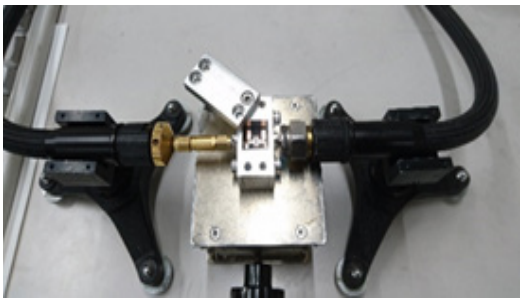
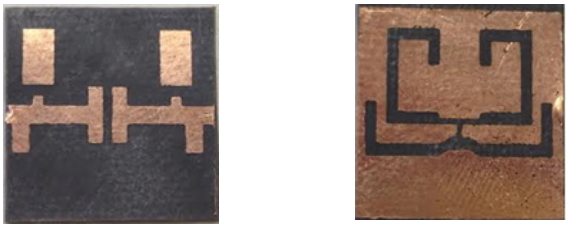


Figure 21. Measurement system.



(a) Patch shape (b) Slot shape

Figure 22. Photograph of the simplified filter.

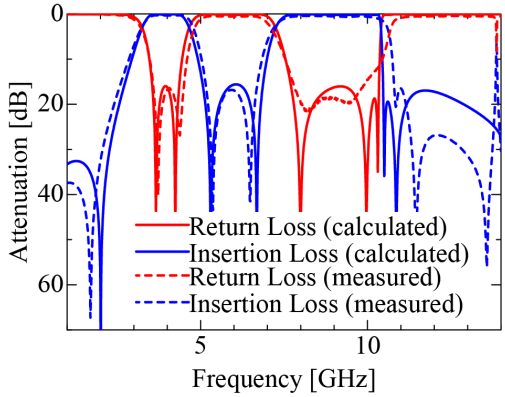
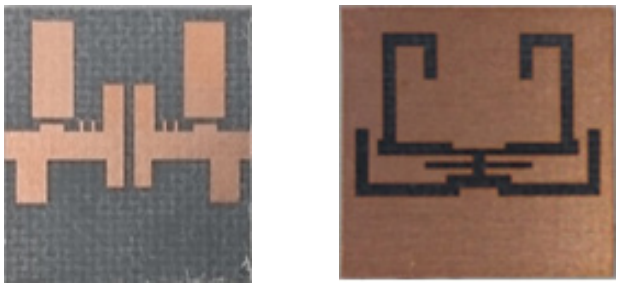


Figure 23. Measurement frequency characteristics of the simplified filter.



(a) Patch shape (b) Slot shape

Figure 24. Photograph of the proposed filter.

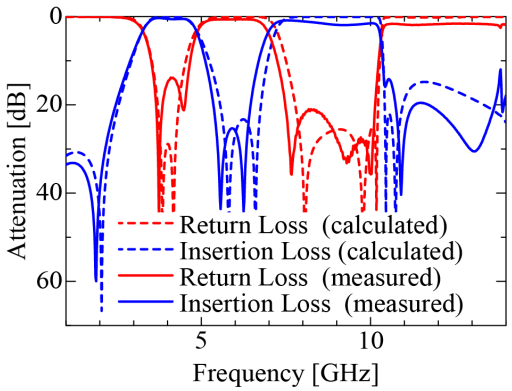


Figure 25. Measured frequency characteristics of the proposed filter.

4. Experimental results

To confirm effectiveness of the proposed filters, we fabricate two kinds of them by the photo-etching technique. As a dielectric substrate, we use Rogers RT/duroid 5880 with dielectric constant ε_r = 2.2 and the thickness t_2 = 0.508 [mm], and also measure the filter characteristics in the shielding box (see the photograph of the measurement system in Fig. 21) by the vector network analyzer (Agilent Technologies Inc. E8361C). The first fabricated filter has the simplified structure shown in Fig. 10 and its photograph is shown in Fig. 22. Comparison of the transmission characteristics between the measured and the calculated results are performed in Fig. 23, where the solid lines indicate the measured results and the dashed lines are the calculated ones. Both results are in good agreement with each other in the insertion and the return losses. The return losses in the two passbands are suppressed to 15 dB, and the insertion loss in stopband is also suppressed to 15 dB. On the other hand, the second fabricated filter has the structure with further improvement of the filter characteristic shown in Fig. 20 and its photograph is shown in Fig. 24. The measured results of the transmission characteristics indicated by the solid lines are compared with the calculated ones of the dashed lines in Fig. 25. The measured return loss in the low-frequency passband is about 15 dB lower that the calculated one, whereas that in the high-frequency passband is more than 20 dB and also the measured insertion loss in the stopband dividing two passbands is more than 25 dB. Although there is some deviation between both results due to the fabrication error for the small patch stubs and the slot stub added to the simplified structure. We can confirm the obvious improvement of the filter characteristics experimentally.

5. Conclusion

In this paper, we have clarified from the electric fields that two passbands in the GA-optimized filter developed by us are constructed by five resonant modes and have proposed the simplified structure based on the resonant mechanism. For further improvement of the simplified filter, the mechanism of the transmission zero is investigated from distributions of the magnetic currents. So that by newly loading both the slot stub and the conductor stub, the return loss more than 28 dB in two passbands and also the insertion loss more than 20 dB in the stopband inbetween passbands have been achieved. Finally, effectiveness of the proposed filter has been verified experimentally. To suppress spurious responses in the frequency region more than 14 GHz is the future work, although we don't investigate it in this paper.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by a Grant-in-aid for Scientific Research(C) (16K06372) from Japan Society for the Promotion of Science.

References

Conference Papers:

[1] Y. Takeuchi, R. Nakano, H. Deguchi, M. Tsuji "Improvement of transmission characteristics in multilayer dual band filter," 2017 IEEE Asia Pacific Microwave Conference (APMC) 13-16, Nov. 2017 https://doi.org/10.1109/APMC.2017.8251528.

[2] T. Kido, H.Deguchi, M.Tsuji "Compact planar bandpass filters with arbitrarily-shaped conductor ptches and slots," IEICE Trans. Electron., vol. E94-C, no. 6 pp.1091-1097, June 2011 https://doi.org/10.1587/transele.E94.C.1091.

[3] L. Zhu, S. Sun, W. Menzel "Ultra-wideband (UWB) bandpass filters using multiple-mode esonator," IEEE Microw. Wireless Compon. Lett., vol. 15, no. 11, pp. 796-798, Nov. 2005 https://doi.org/10.1109/LMWC.2005.859011.

[4] R. Li, L. Zhu, "Compact UWB bandpass filter using stub-loaded multiple-mode resonator," IEEE Microw. Wireless Compon. Lett., vol. 17, no. 1, pp. 40-42, Jan. 2007 https://doi.org/10.1109/LMWC.2006.887251.

[5] L. Han, K. Wu, X.P. Chen, "Compact ultra-wideband bandpass filter using stub-loaded resonator," Electron. Lett., vol. 45, no. 10, pp. 504-506, May 2009 https://doi.org/10.1049/el.2009.0510.

[6] Y. Zhou, B. Yao, Q. Cao, H. Deng, X. He, "Compact UWB bandpass filter using ring open stub loaded multiple-mode resonator," Electron. Lett., vol. 45, no. 11, pp. 554-556, May 2009 https://doi.org/ 10.1049/el.2009.3724.

[7] H. W. Deng Y. J. Zhao L. Zhang X. S. Zhang S. P. Gao "Compact quintuple-mode stub-loaded resonator and UWB filter," Progress In Electromagnetics Research Letters., vol14, no. 8, pp. 111-117, 2010 https://doi.org/10.2528/PIERL10030912.

[8] K. Song, Q. Xue, "Novel broadband bandpass filters using Y-shaped dual-mode microstrip resonators," IEEE Microw. Wireless Compon. Lett., vol. 19, no. 9, pp. 548-590, Sept. 2009 https://doi.org/10.1109/LMWC.2009.2027058.

[9] L. Han, K. Wu, X. Zhang, "Development of packaged ultra-wideband bandpass filters" IEEE Trans. Microw. Theory Tech., vol. 58, no. 1, pp. 220-228, Jan. 2010 https://doi.org/10.1109/TMTT.2009.2036399.

[10] Q.-X. Chu, X.-K. Tian, "Design of UWB bandpass filter using stepped-impedance stub-loaded resonator," IEEE Microw. Wireless Compon. Lett., vol. 20, no. 9, pp. 501-503, Sept. 2010 https://doi.org/10.1109/LMWC.2010.2053024.

[11] J. Xu, W. Wu, W. Kang, C. Miao, "Compact UWB bandpass filter with a notched band using radial stub loaded resonator," IEEE Microw. Wirel. Compon. Lett,. vol. 22, no. 7, pp. 351-353, July 2012 https://doi.org/10.1109/LMWC.2012.2201930.

[12] Z.X. Zhang, F. Xiao, "An UWB bandpass filter based on a novel type of multi-mode resonator," IEEE Microw. Wirel. Compon. Lett., vol. 22, no. 10, pp. 506-508, Oct. 2012 https://doi.org/10.1109/LMWC.2012.2218229.

[13] H. Zhu, Q.-X. Chu, "Compact ultra-wideband (UWB) bandpass filter using dual-stub-loaded resonator (DSLR)," IEEE Microw. Wireless Compon. Lett., vol. 23, no. 10, pp. 527-529, Oct. 2013 https://doi.org/10.1109/LMWC.2013.2278278.

[14] L.C. Lin, S. Yang, S.J. Sun, C.H. Liang, "Ultra-wideband bandpass filter using multi-stub-loaded ring resonator," Electron. Lett., vol. 50, no. 17, pp. 1218-1220, Aug. 2014 https://doi.org/10.1049/el.2014.1256.

[15] X. Li, X. Ji, "Novel compact UWB bandpass filters design with cross-coupling between short circuited stubs," IEEE Microw. Wirel. Compon. Lett., vol. 24, no. 1, pp. 23-25, Jan. 2014 https://doi.org/10.1109/LMWC.2013.2287231.

[16] A. Taibi, M. Trabelsi, A. Slimane, M. T. Belaroussi, J. P. Raskin, "A novel design method for compact UWB bandpass filters," IEEE Microw. Wireless

Compon. Lett., vol. 25, no. 1, pp. 4-6, Jan. 2015 https://doi.org/10.1109/LMWC.2014.2363016.

[17] S. Kumar, R. D. Gupta, M. S. Parihar, "Multiple band notched filter using C-shaped and E-shaped resonator for UWB applications," IEEE Microw. Wireless Compon. Lett., vol. 26, no. 5, pp. 340-342, May 2016 https://doi.org/10.1109/LMWC.2016.2549700.

[18] M. Tsuji, H. Deguchi, A. Kido, M. Ohira "Multi-resonator generation by genetic optimization for application to panar-circuit bandpass filters," Special issue on recent progress in computational electro magnetics and its applications. Volume 129 (2009) Issue 10 Pages 681-686 https://doi.org/10.1541/ieejfms.129.681

[19] A. Kido, H. Deguchi, M. Tsuji, M. Ohira "Multi-resonator generation in arbitrarily-shaped planar-circuit filters by genetic optimization," 2007 European Microwave Integrated Circuit Conference 8-10 Oct. 2007 https://doi.org/10.1109/EMICC.2007.4412744

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 21-25 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Scientific Principles behind 2017 World Universiade of Safety Lantern and Torch

Wei Long Chen, Fang Lin Chao*

Department of Industrial Design Chaoyang University of Technology,436, Taiwan R.O.C

ARTICLE INFO

Article history:
Received: 11 May, 2018
Accepted: 12 June, 2018
Online: 02 July, 2018

Keywords:
Scientific principles
Safety lantern
Torch
Convection

ABSTRACT

The torch relay for the 2017 Taipei Summer Universiade started June. The innovative torch design was altered and tested, follow-up production successfully. For preventing weather effect, the head of the torch embedded with a unique airflow structure that allowed the sacred flame to pass the wind and rain tests. The safety lantern displays a 3D spiral flash. Safety lantern uses biofuel so that the mother flame can transport by air. The scientific principles are Stack effect, Coanda effect, Bernoulli's principle, and Capillary phenomenon. Product design refers to principles and a series of work from the specifications to determine the product structure. Cross-discipline collaboration is needed to increase the integrity.

1. Introduction

The torch relay for the 2017 Taipei Summer Universiade started June 20 in the Italian city of Turin [1], "with a vision to shape the leaders of tomorrow through their experiences of international student sport." It was then passed by Professor Gianmaria Ajani, the rector of the University of Turin, to Matytsin. He then transferred the flame to the Taipei 2017 Torch, which was held by Ko Wen-je, the Mayor of Taipei (Figure 1). The domestic leg of the Torch Relay took the flame and visited 22 cities and counties.

The unique design developed by cooperation of both designers and engineers. The torch design included Taiwan's essential design elements and craft skills. With the collaboration of scientist and designer, the concept was revised and tested, follow-up production successfully [2]. For preventing weather's affection, the head of the torch embedded with a unique airflow structure that allowed the sacred flame to pass the wind and rain tests. The safety lantern displays a single three-dimensional spiral flash (Figure 2).

Safety lantern have several vertical vanes, each vane has a section encircled by the concealing member defining a flow accelerating zone. External air flows into the flow diverting mechanism through the air intake ports, circumferentially conceals with the wall.

Figure 2 also shows the concept of torch design comes from Confucian's article "Gentlemanly fair play." The Master said, "Supportive, encouraging, congenial – such a man may be called a gentleman." Fair play induces supportive and encouraging with a player, congenial with friends.

The torch is made of metal, on which there are an array of ventilating holes. The lower holding portion made of weaving bamboo comes from eastern civilization.

The safety lantern keeps the mother flame continuously so that the fire can pass through the different location. The design includes a safety feature to avoid accidental burning if the fuel leaks out. There is no CO generated during combustion, so the safety lantern can transport using commercial air traffic.



Figure 1. Torch used in 2017 Taipei Summer Universiade [2]

*Corresponding Author: Fang Lin CHAO, Email: flin@cyut.edu.tw



Figure 2. The concept of torch design comes from Confucian [2], design team: H.L Chang of Ui-D Taiwan. The safety lantern keeps the mother flame continuously, the torch table presenting the flash at the main venue of the sports competition, and the relay torch. The height of safety lantern and torch are 50 cm and 70cm.

2. Scientific Principles

The spiral frame of safety lantern is distinct and has a stable shape and an elongated height (Figure 2). Air driven by buoyancy goes upward and draws in fresh air through the passages which make air flowing therein rotate and go upward.

For the torch, after gas and air mixing process and is being ignited by the igniting needle, combustion flame flowing out from the top cap and fresh air drawn from the side holes for mixing. Gas and air mixture undergoes the mixing processes and flows out of the plurality of exit.

2.1. Stack effect

The safety lantern has a concentrated hot zone around and above the combustion head due to combustion flame and spiral flow around. Heated air with lower density in the heated region flows upward and creates low pressure to draw fresh air from intake port into the transparent shield, thereby increasing the height of flame in the torch. This phenomenon is known as stack effect or chimney effect.

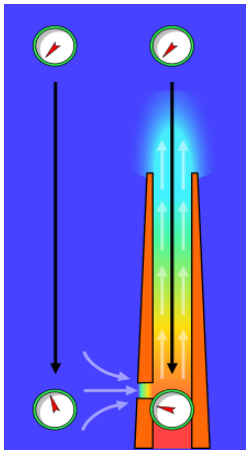


Figure 3. The safety lantern: stack effect [3] (gauges represent absolute air pressure, airflow indicates with grey arrows)

2.2. Coanda effect and centrifugal force effect

The passages direct fresh air toward the inner surface of a transparent shield at a specific angle to create a spiral air flow pattern and tangent to the inner surface of the shield. This feature dramatically helps to stabilize swirling air flow pattern due to Coanda effect and centrifugal force.

Coanda effect [4] related to the tendency of a fluid jet to stay attached to a convex surface. The pressure difference across the jet causes the jet to deviate towards the nearby surface Figure 4, and then to adhere to it. The separation angle is a function of the Reynolds number of the fluid.

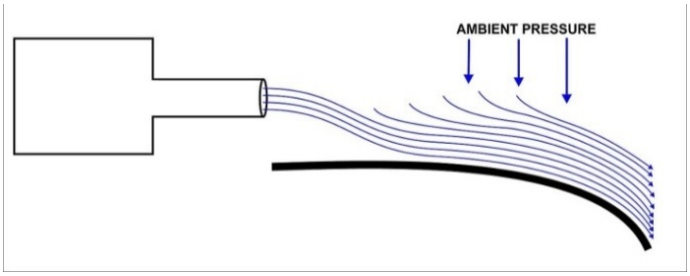


Figure 4. The jet will cling to the surface because any molecules of ambient air that move into the space between the jet and swept away, adhere to the surface even if it is curved [4].

The fluid stream attaches a contour when the flow is directed at a tangent to that surface. This spiral motion of air flow also induces centrifugal force. The safety lantern takes advantages of stack effect and Coanda effect which help to create a stable spiral and elongated flame. The centrifugal force is an inertial force directed away from the axis of rotation that appears to act on objects when viewed in a rotating reference axis [5]. In a rotating reference frame, objects appear to be under the influence of a radially outward force that is proportional to their mass, to the distance from the axis, and to the square of the angular velocity [4]. The detail of structure in the safety lantern shows in Figure 5.

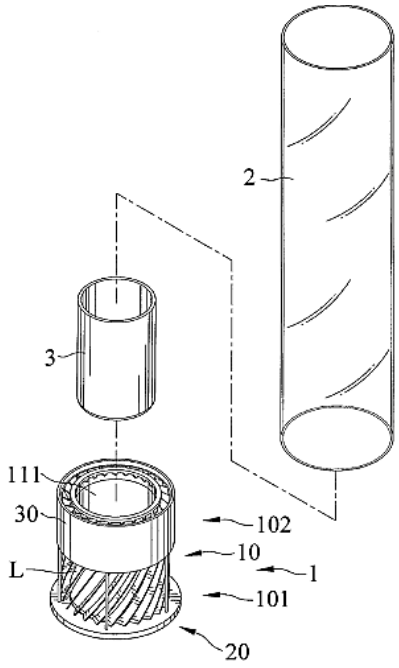


Figure 5. The safety lantern [7]: a perspective view of the flame device, where 1: assembly, 2: glass tube, 3: container, 20: basement, 30: cover guider, 10: passages, 101: inlet end, and 102: acceleration zone.

The safety lantern produces a swirling flame, and a stable fluid creates within the guided shield channels. Heated air in the shield creates buoyancy due to a difference in air density. The perspective view (Figure 5) showing the flame device producing a stable and prolonged flame in a steady fluid field. Undergoes convection directs air through the transparent glass shield (#2). Centrifugal forces also keep air moving spirally and drive air into the safety lantern.

2.3. Capillary phenomenon

The capillary phenomenon is the ability of a liquid to flow in narrow spaces in opposition to gravity [6]. The effect can find in the drawing up of water between the hairs of a paint-brush. If the diameter of the tube is sufficiently small, then the combination of surface tension and adhesive forces dominate. The thinner the space in which the liquid can travel, the further up it goes. In torch design, we choose liquid transparency "biofuel" as fuel.

The capillary phenomenon presents while the biofuel's viscosity is low. This property allows the fuel transport from the fuel tank to the upper ignition area. When fuel is filled into the body below the wick, the wick can volatilize fuel for combustion. The cord placed in the holding space and the two jaws can hold the wick so that the wick's head end located in the burn area [7]. The user can turn to adjust the elastic, change the clamping gap when clamping the cord will be compressed. Adjust the speed at which the wick draws fuel, thus controlling the size of the combustion flame. The Use of metal mash wick also removes the disadvantage of waste and burn out of traditional cotton thread core.

2.4. Bernoulli's principle

Bernoulli's principle states that an increase in the speed of fluid co-occurs with a decrease in pressure. Bernoulli's equation is sometimes valid for the flow of gases: provided that there is no transfer of kinetic or potential energy from the gas flow to the compression or expansion of the gas. Bernoulli equation applies to incompressible fluids, and steady compressible fluids up to approximately Mach number 0.3 [8]. There are many applications in surrounding life, such as an injector on a steam locomotive. The pitot tube is used to determine the airspeed. The flow speed of fluid can also be measured using the Venturi meter [9]. The pressure in the reduced diameter region decreases and display in the meter by liquid height difference.

The main body of the torch is an aluminum metal rod. The hollow metal tube contains the gas supply lines. The high-pressure gas is conveyed upward by a metal pipe from a fuel tank in below. The upper section has an enlarged gas mixing zone. At the top of the gas mixing zone is the metal flame exit cover. The high-pressure gas gradually spreads to the top metal cover of the torch. As shown in Figure 6, the central hole on the metal cover is the exit of the main flame. There are four big holes next to it, which is the side flame exit. When the gas in the torch ejectes at high speed, the pressure in the central area becomes smaller; therefore, the external air will flow from the side of the leaf-shaped opening into the mixed gas area (Figure 7). When the outside air enters the mixed gas zone, it can also cool the aluminum rod; it avoids the chance of being exposed to the outside of the torch.

The surrounding metal shell wrapped on the outside of the aluminum metal tube. The hand grip is under the metal shell, and the grip is a structure woven with bamboo. Burning the flame increases the temperature of the aluminum metal tube. There is an air layer between them (Figure 7), the thermal conductivity of the air is very low which reduced the amount of heat transfer between internal pipe and the periphery. When cold air enters from the intake holes, it also cools the metal pipes at the same time.

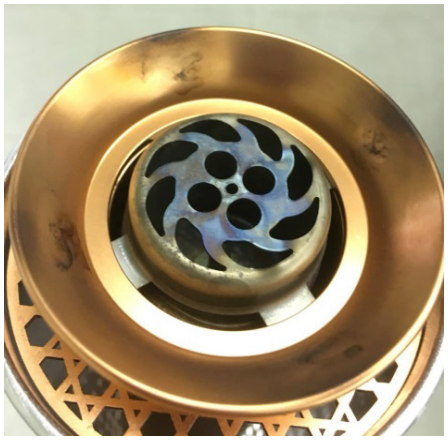


Figure 6. The metal cover on top of the torch: The central hole is the exit of the main flame. There is four side flame exit. The outer leaf-shaped opening is air inlet of the mixing zone.

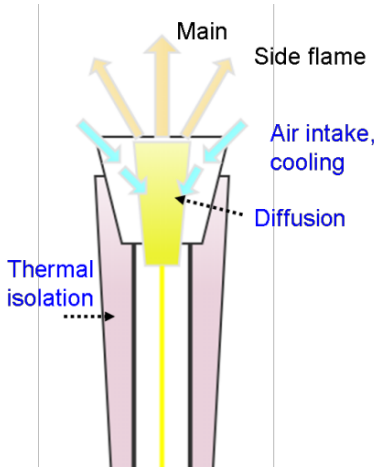


Figure 7. Torch: External airflow driven by Bernoulli's principle, where the gas ejects and spreads (diffusion). The low pressure in the central area causes the external air enters the air intake holes into the mixing zone. The air between the aluminum rod and surrounding metal decoration also provide thermal isolation.

3. Experimental Evaluation

In the previous chapter, we discussed the scientific principles behind the design of the torch. In this chapter, we utilize two prototypes and perform functional tests. The purpose of the experiment is verifying the benefits of the above scientific principles in design. The prototype of the product is one-to-one scale and tested according to the conditions of use. In addition to the shape of the flame, we also care about the safety of the user, and we hope to appropriately lower the temperature of the product in areas that may be in contact with user's body.

3.1. Safety Lantern

Figure 8 indicates the evaluation of the experimental prototype. We observed safety lantern with spiral shape and an elongated height. As the air is heated, it rises along the guide channel, and the airflow is a steady laminar flow. Because of the Coanda effect, the airflow is attached to the inner wall of the glass and rotates upwards. As the stack effect and centrifugal force continue to rise

upward, a stable flash formed in the lower half zone. After passing through the middle area, transition and turbulent zone take place, and the shape of the flame is irregular.

The combination of the mentioned effects can substantially elongate the height of flame and change the swirling speed of flash according to flame visualization required. If the flame guiding member is absent, it is difficult to control a smooth flow. The flame guiding member allowed controlling natural convection and drawn fresh air from outside the chamber. The flame guiding member of 44-degree skew angle shows a reasonable display quality (Figure 8). When cold air is introduced, it also cools the metal and glass tube. The temperature measurement of aluminum castings and iron plate indicated the contact temperature is not higher than 60 degrees C.



Figure 8. Evaluation of experimental prototype. There are three flame zones; laminar zone (lower), transition zone (middle), and turbulent zone (upper).

3.2. Torch

There are five flame outlets above the fire exit, which scatter at different angles respectively to show the flame shown in Figure 9 and 1. The runner at the fuel outlet must be balanced so that the flame maintains a similar height in all directions. It successfully demonstrated the five lights which symbolizing the Olympic Games.



Figure 9. The flow diverting mechanism design of the torch

At the previous trial, the torch flame that burns on mixing gas from first mixing process undergoes convection directly with air, thereby creating a more significant portion of colorless and transparent flame, causing a less visible swirling pattern and shorter in flame height. We will expand the gas mixing area at the gas outlet, let the gas mix and send it out from the top hole. After gas and air mixed in the chamber, the torch shows a reasonable display quality.

4. Cross-discipline Collaboration

Although scientific principles are well known, designer seldom can integrate. Product design refers to a series of work from the specifications to determine the product structure. The designer must take into account the form, technology, and regulations, through the appropriate product form to reveal differentiated value.

4.1. Integration

We experienced cross-discipline collaboration is vital to torch design. It covers material, combustion, fluid dynamics, heat transfer, usability, and aesthetics. One of successful field practice project is EU-IDES-EDU: "Master and Post-Graduate education and training in multi-disciplinary teams." In [10], describes the necessity of more integrated and cross-disciplinary approaches to building design. The challenges encountered during implementation of the design education across different professions were discussed. The Integrated Design Process (IDP) was developed to streamline the design process of solar integrated low-energy buildings. The involvement of engineers in early stage improved design significantly.

We also experienced idea jumping iteratively through the stimulus of CAE simulation and prototype experiment during torch design. In [11] authors present "speculative design as a means to technological fluency." Speculate means form a theory or conjecture about a subject without firm evidence. Creativity's natural imply explore and discover a new connection between facts or knowledge. Firm evidence is only a stepping stone to problem-solving. Technological fluency is a term used to describe a level of expertise: the capability to understand, use, and assess technology beyond its rote application. By using a Speculative Design approach to extend the idea of technological fluency in new directions, broadening understandings of the development of technology and adding a dimension of criticality [12].

Design often plays a central role in technology fluency programs. The design process typically derived from engineering approach in which functional requirements drive the process toward the solution. Speculative design project involves the construction of prototypes. The education of the designer's technological imagination consists in learning how to use tools, expression, techniques for rapid prototyping, processes of iterative experimentation, and skills in social negotiation and integration [12].

Cross-disciplinary and integrative are significant attributes of technological fluency in speculative design projects. Pressures internal and external to engineering courses for the adoption of multidisciplinary team-based work are discussed. Some key factors arising from the survey are addressed [13].

4.2. The design process and CAE

The critical elements of renovation in torch design are related to cross-discipline and knowledge. As shown in Figure 10, the design guide help prevent mistake arrangements. CAD/ CAE are used to realize the idea quickly.

There are many engineering issues in torch design; such as heat transfer, fluid dynamic, mechanical stress and other physical principles. The flow field simulation enables fine tune of geometric design. Functional test proof and confirm the product features, through a series of measurements such as thermal energy, flame height, contact temperature (glass and metal). After that, the checklist and regulations were reviewed to enable the modified concepts.

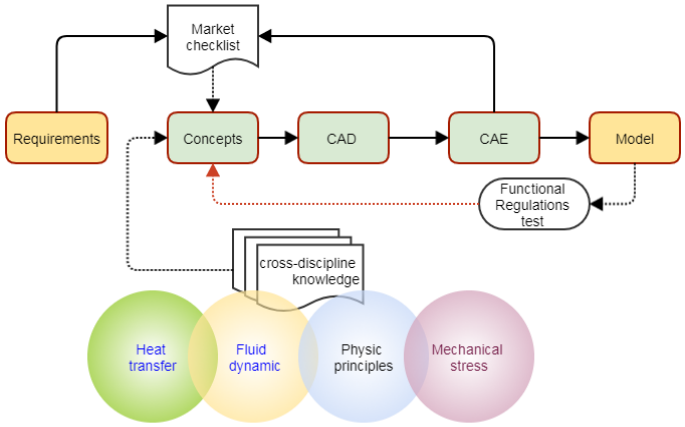


Figure 10. Design process

We built the digital model with CATIA. As shown in Figure 10, the prototype model can utilize for measuring actual physical parameters. Comparing the measured data and simulation results, we refine the CAE models. When the settings are confirmed, the designer can use the digital model to obtain design parameters. After that, a functional test proof and validate the product features. The cross-discipline knowledge help prevent contradiction design. The CAE is used to verify the idea quickly and shortens development cycle time.

The role of a designer is changing according to the evolutional technology and society. From the torch project, product renovation needs to cooperate with others expertise closely. Understanding the relevant principles in the product can lead to support through cross-domain collaboration.

5. Conclusions

In this article, we discussed the scientific principles within the torch and safety lantern with spiral shape and an elongated height; namely Stack effect, Coanda effect, Bernoulli's principle, and Capillary phenomenon. After gas and air mixing process and is being ignited, combustion flame flowing out from the grooves and fresh air drawn from the passages mix. Air in the shield driven by buoyancy goes upward and brings in fresh air through the passages which make air flowing therein rotate and go uphill.

The process of design also provided. We employee qualitative and CAE design process to fulfills its primary and auxiliary functions. Design process helps understand combustion phenomenon and trend. Through the model test, designer and engineers can select proper parameters. Cross-discipline collaboration is needed to increase the integrity of the concept. Designers should experience the scientific principles to collaborate with other team members in cross-domain development.

Acknowledgments

We wish to thank design effort of Ui-D, and students in CAE Lab in Chaoyang University of Technology for supplying the testing and measurement. This research partially supported by the Pro-Iroda company with prototyping.

References

[1] https://www.insidethegames.biz/articles/1051799/taipei-2017-torch-relay-begins-in-turin

[2] https://udn.com/news/story/11406/2671767

[3] https://en.wikipedia.org/wiki/Stack_effect

[4] https://en.wikipedia.org/wiki/Coand%C4%83_effect

[5] https://en.wikipedia.org/wiki/Centrifugal_force

[6] https://en.wikipedia.org/wiki/Capillary_action#Phenomena_and_physics_of_capillary_action

[7] Chen, Wei-Long, and Jan-Nan Chen. "Device for producing stable and augmented flame." U.S. Patent 8,641,413, issued February 4, 2014. https://patents.google.com/patent/US8641413B2/en

[8] White, Frank M. Fluid Mechanics, 6th ed. McGraw-Hill International Edition. p. 602, 2017.

[9] Kamela, M., "Thinking about Bernoulli," The Physics Teacher, **45**(6), 379-381, 2007. https://doi.org/10.1119/1.2768700

[10] C., Brunsgaard, P. Dvořáková, A. Wyckmans, M. Laskari, "Integrated energy design–Education and training in cross-disciplinary teams implementing energy performance of buildings directive (EPBD)", Building and Environment, **72**, 1-14, 2014. https://doi.org/10.1016/j.buildenv.2013.10.011

[11] Lukens, Jonathan, and Carl DiSalvo. "Speculative design and technological fluency." International Journal of Learning and Media, **3**(4), 23-40, 2011. doi:10.1162/IJLM_a_00080

[12] Burdick, Anne, and Holly Willis. "Digital learning, digital scholarship and design thinking." Design Studies, **32**(6), 546-556, 2011. https://doi.org/10.1016/j.destud.2011.07.005

[13] Denton, H. G. (1997). Multidisciplinary team-based project work: planning factors. Design Studies, **18**(2), 155-170, 1997. https://doi.org/10.1016/S0142-694X(97)85458-0

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No 5, 23-27 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

An Optical-based Fingertip Force Sensor

Emanuele Lindo Secco*, Taye F. Agidew, Atulya Kumar Nagar

Robotics Laboratory, Department of Mathematics and Computer Science, Liverpool Hope University

ARTICLE INFO

Article history:
Received: 20 July, 2018
Accepted: 14 August, 2018
Online: 08 September, 2018

Keywords:
Bio-mimetics
Fiber-optics
Force sensing
Optical sensing

ABSTRACT

This work introduces the design and prototype development of a novel optical-based fingertip force sensor, which is integrated in a bio-mimetic finger for robotic and prosthetic manipulation. Given the optical nature of the sensor, the proposed device is free of any electrical and metal components. Accordingly, the design improves manipulation of objects with high electromagnetic compatibility performance, reducing sensor's susceptibility in the presence of these disturbances. The sensor has inherently high SNR and low-power consumption. The concept of using simplified and low cost optical based fingertip force sensor for bio-mimetic anthropomorphic fingers has not been widely implemented so far. The sensor is based on a cantilever design combined with a couple of fiber optics and a reflective surface, it is integrated on the distal phalange of a novel bio-mimetic and anthropomorphic robotic finger. Sensor design was performed and optimized throughout a 3D printing process and Finite Element Analysis. The sensor's sensitivity (0.098V/N), resolution (0.01N), accuracy (3% of full scale output (FS)) and hysteresis (9.24% of FS) were characterized through a calibration processes. Various applications like surgical manipulation or handling in harsh environment (i.e. high voltage, high temperature and explosive environments) will benefit from the proposed sensor's performance, reliability, simplicity and bio-compatibility.

1. Introduction

Many multi-fingered robotic hands with multi-axial force and torque sensors have been developed. These multi axial force sensors usually employ strain gauges driven by electrical circuitry and are made of metal components [1]. The use of sensors with electrical circuit and metallic components increases the sensor's vulnerability to electrical or magnetic fields [2]. In contrast with multi-axial force sensors based on strain gauge technology, sensors designed based on the use of fibre-optics has several benefits including simplified design, immune to electromagnetic fields, miniaturization, low-level noise, has low power consumption and no need for any electronic noise filtering [3]. In this context, previous research has already proposed interesting applications on the development of force and tactile sensors with the use of metal free components and biologically friendly solutions which, moreover, will also guarantee compatibility in medical scenarios [2-3].

This work has been proposed due to the following limitation observed from current fingertip force measuring systems:

- Complex tactile system design
- High electromagnetic vulnerability
- Unavailability of simplified and low cost optical based fingertip force sensor design for biomimetic anthropomorphic fingers.

The concept of using simplified and low cost optical based fingertip force sensor for biomimetic anthropomorphic fingers has not been widely implemented so far. The proposed fingertip force sensor takes advantage of the use of fibre-optics and is integrated in a biomimetic finger in order to get tactile information during grasping and manipulation of objects. The objective of this work is designing optical based fingertip force sensor. The design involves only the distal phalange of the finger. The phalange has the same shape, arrangement, length and proportion as its human counterpart [4].

2. Sensor Design

To exactly imitate the size and shape of the human index finger bones, the laser-scan model of human finger was used. The Stereo Lithography (STL) format of the model was first imported into

*Corresponding Author Emanuele Lindo Secco, seccoe@hope.ac.uk

Autodesk123 (Autodesk, Inc.). In this stage the original file was further modified by introducing a cantilever beam design, holes for placement of optical fibre and grove between sensing head and reflective surface (Figure 1).The design involves optical fibre, mirror and built-in cantilever beam [2]. After modifying the file, it was further analysed in Autodesk Mesh mixer to reduce the number of meshes [5] in the model. The aim of mesh reduction was to easily simulate Finite Element Analysis (FEA) of the model in Solid Works (Dassault Systèmes Solidworks Corp) [6]. Finally the model with reduced mesh (Figure 2) was exported to Solid Works and the FEM analysis was performed. The standard parameters of phalanx of index finger [7] are shown in Table 1.

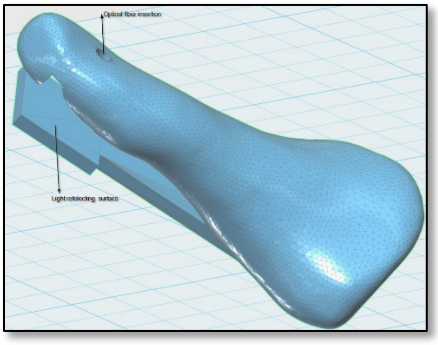


Figure 1: Built-in cantilever beam and optical fibre holes: the wired fibre will be inserted through the hole in order to design the tactile sensor (see also Figures 4 and 5)

Table 1: Parameters of the index bio-mimetic finger: length and weight of the index finger phalanges

Index finger phalanx	Length(mm)	weight(g)
MCP to PIP	53.4	5.5
PIP to DIP	32.0	2.0
Distal phalanx	23.7	1.2

2.1. FEA Simulation

The purpose of FEA simulation is to ensure how much amount of force the fingertip can resist without undergoing plastic deformation. First geometry or shape of finger phalanx was defined. Next, an appropriate material (i.e. Acrylonitrile Butadiene Styrene, ABS) for the model has been selected. Mechanical properties [8] of ABS have been used as input for the FEA simulation (Table 2).

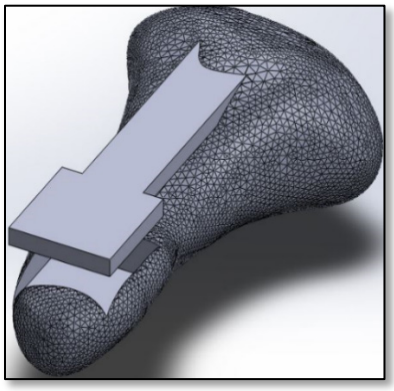


Figure 2: Reduced mesh of the distal phalanx of the index finger: the mesh was distributed along the fingertip in order to perform the FEM

Table 2: Mechanical Properties of the ABS material

Property	Value	Units
Elastic Modulus	2000	N/mm^2
Poisson's Ratio	0.35	
Shear Modulus	318.9	N/mm^2
Mass Density	1020	Kg/m^3
Tensile Strength	40	N/mm^2
Compressive Strength	42	N/mm^2

The mean maximum external finger force (i.e. 5[N]) for static prehension is applied to the phalanx of index finger [9]. Finally FEA analysis of the 3D model been performed. The description of how the applied forces spread throughout the phalanx is illustrated in Figure 3.

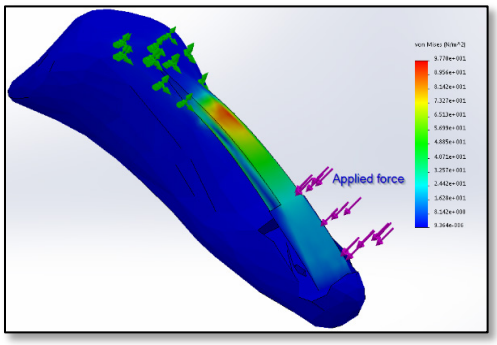


Figure 3: FEM simulation of the cantilever (ABS material, load of 5[N]): force is applied to the cantilever in order to estimate Von Mises stress

The red portion of the simulation indicates high stress concentration areas and the blue part shows unloaded or stress free region of the phalanx Figure 3. According to the finite element simulation result the maximum stress applied on the beam due to the applied force is much lower than, the yield strength of ABS. This means the applied force doesn't produce mechanical failure or damage to the built in cantilever beam of the phalanx and it shows the design is safe.

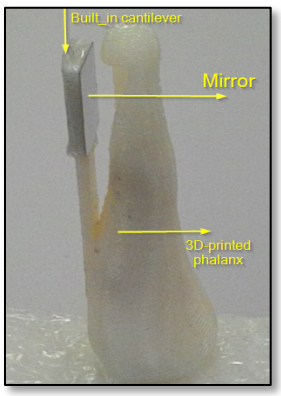


Figure 4: 3D printed distal phalanx of index finger: the mirror allows measuring the cantilever flexion according to the applied load and therefore to transduce such a flexion into a modulation of the intensity light provided through the fibre (see also Figures 1 and 5)

2.2. 3D Printing and sensor component integration

After the FEM simulation, the real distal phalanx of index finger was fabricated using HP 3D printer. The phalanx has built-

in cantilever beam used for positioning light reflective surface. The movement of this beam moves the mirror and modulates the intensity of reflected light back to the receiving terminal of the optical fibre. The 3D printed phalanx is shown in the Figure 4.

2.3. Working principle of the sensor

The light beam is projected on a moving cantilever beam reflective surface by fibre optic cable (i.e. the emitter), a fraction of the reflected beam is received by the second optic fibre (i.e. the receiver) that are closely located [1] (Figure 5). The reflected light intensity is a function of the amount of force applied at the fingertip. The applied force changes the distance between the reflective surface and the sensing head of optical fibre cables. This change modulates the intensity of reflected light and produces different voltage outputs. The amount of output voltage from the sensor is directly proportional with the force applied at tip of the phalanx.

3. Sensor Calibration

To use the designed sensor, the sensor has to be calibrated. The calibration should convert force applied to the fingertip into corresponding voltage [2].

3.1. Calibration set-up

The experimental consists of a Keyence FS-N11MN Digital Fibre Optic Amplifier, IIT-FT17 force (IIT-FT17 force & torque sensor), which is combined with a torque (F/T) sensor: sensor data are acquired via an NI USB-6000 Data AcQuisition (DAQ) board (National Instruments Inc.) while the sensor is mounted on a mechanical vice as it is reported in Figure 6. The phalanx of the finger is fixed on the vice, while the force sensor is physically connected with the moving element of the vice. This solution allows to manually modulate the effective applied force to the phalanx, i.e. to the sensor. During the experiments, the electrical output of the FS-N11MN amplifier is sent to an analogue pin of the DAQ. Then, a LabView (National Instrument, NI) graphical interface captures the output signal in real-time, while another program – developed in C++ Programming Language – acquires the force sensor outputs. This further measurements are processed via another LabView mask. These data (i.e. output of the two force sensors) are saved and processed.

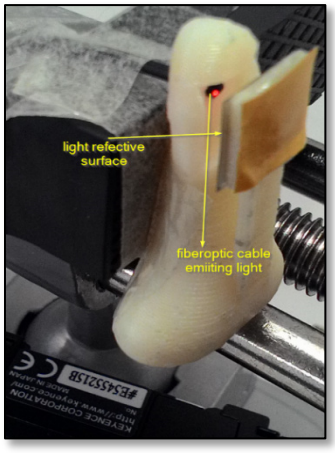


Figure 5: Fingertip fibre optic cable and mirror arrangement (see also Figures 1 and 4)

3.2. Calibration procedure

Different trials were performed while applying force to the sensor and releasing the load, thanks to the linear guide which was provided by the vice in Figure6. The IIT-FT17 force and torque (F/T) sensor was used to compare the reading vs. the fibre-optic sensor and therefore calibrate this latter one. Different values of force were applied. Precisely, a sequence of 15 values of forces were applied during the 'loading' and 'unloading' stages. The NI DAQ board was used to acquire and save the value of the signals (i.e. the voltage) from the KEYENCE amplifier and the reference F/T sensors. Signals were acquired every 2 ms, namely at a sampling frequency of 500 Hz.

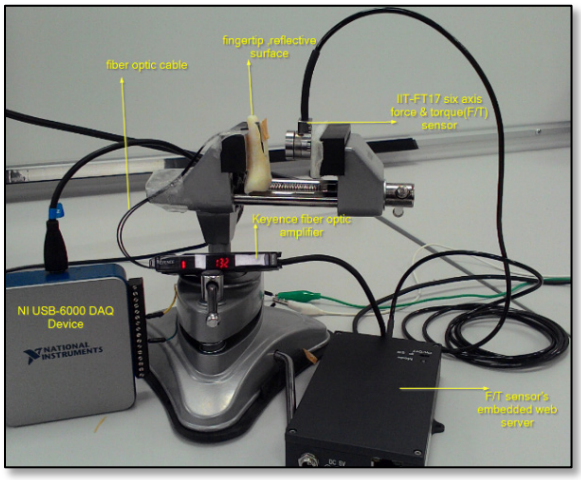


Figure 6: Force sensor calibration setup: load is applied to the fingertip and sensor output is measured accordingly (see the details on Paragraph 3)

4. Sensor Performance

4.1. Regression analysis

Following the experiment and data acquisition, all information were re-shaped into a data base format to be analysed and to identify a proper calibration between the applied load and the voltage output of the sensor. According to a preliminary qualitative analysis of the data, linear regression method was adopted in order to find a proper fitting of the data [10]. The electrical signal was correlated with the applied force on the finger, finally, a fitting curve was obtained (Figure 7). Here, the linear regression equations are reported:

$$V = 0.098 \cdot F + 1.10 \quad (1)$$

$$F = 10.22 \cdot V - 11.29 \quad (2)$$

where V and F are representing the sensor signal and applied force, respectively.

Thanks to the Equation (2), the effective force can be calculated, provided that the output voltage of the sensor is measured. Moreover, these equations define the sensor sensitivity, namely 10.22 N/V [8]. A regression analysis shows that the relationship between the force and the output of the sensor is quite linear with a coefficient of determination ($R^2 = 0.9739$): more than 97.39 % of the electric voltage is due to the load which is applied to the fingertip. An analysis on the standard deviation of the

prediction of the value of the force provides a result of 1.86%: a small deviation occurs between the data vs. the actual values.

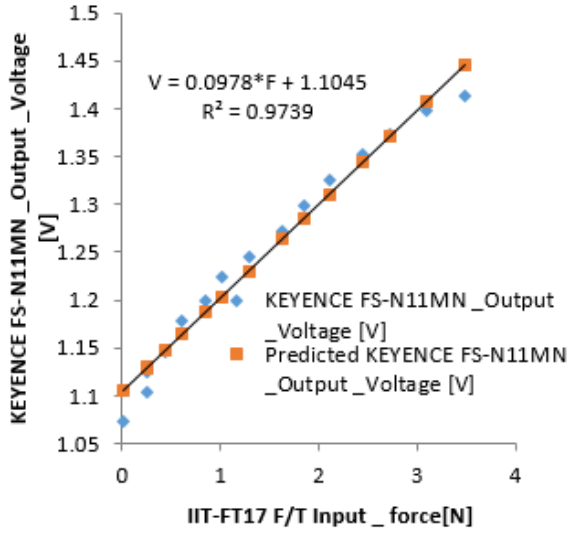


Figure 7: Calibration curve of force sensor performed with a regression analysis (see details in Paragraph 4)

4.2. Sensor accuracy

In this context accuracy is measure of closeness between actual sensor readings and its corresponding predicted value. The sensors accuracy is measured by its relative and absolute errors. The sensor relative error is the difference between the actual voltage reading obtained from KEYENCE FS-N11MN Fibre Optic Sensor and the corresponding predicted voltage value obtained from fitted linear equation during the calibration procedure. Table 4 shows the actual and predicted voltage values. Based on the data given on Table 4 percentage of sensor's maximum relative error is 3.1097%. This means the sensor introduces maximum of ±0.03109 error from each of its readings and is indication of sensors better accuracy.

Table 3: Results of the Regression Analysis which brought to the linear calibration expressions (see details in Paragraph 4)

Parameter	Value
Multiple R	0.9868
R^2	0.9739
Adjusted R^2	0.9719
Standard Error	0.0184
Observations	15

4.3. Sensor range & resolution

The designed sensor works over an input range between 0[N] and 4[N]. If exceeded the specified design range ,it result in permanent damage or destruction of sensor. The corresponding output range of the sensor is between 1.1045 [V]and 1.6[V].

The smallest input force value the senor can detect [11] is 0.01[N]. In other words the sensor can discriminate input force values above 0.01[N].

4.4. Sensor sensitivity & saturation

It is change in output voltage of the designed sensor per unit change in input force. The slope of fitted linear regression equation (i.e. Equation 1 is sensitivity [11] of the sensor (i.e. 0.0978 V/N or 10.2249 N/V).

Table 4: Actual and predicted reading of the voltage outputs

Sensor's actual reading	Predicted value	Residual
1.070	1.110	-0.033
1.100	1.120	-0.250
1.120	1.130	-0.005
1.145	1.147	-0.002
1.170	1.160	0.013
1.198	1.186	0.012
1.224	1.203	0.021
1.245	1.230	0.015
1.272	1.2638	0.009
1.298	1.285	0.013
1.325	1.310	0.014
1.351	1.343	0.007
1.373	1.370	0.002
1.3973	1.406	-0.009
1.413	1.446	-0.033

As the input force applied to the fingertip increases the distance between optical fibre and reflective sensor decreases and sensor's output voltage increases. Beyond its input range (i.e. 4[N]) the sensor reaches its saturation limit. This means, any further increment of input force will no longer increase [11] output voltage. Sensor's saturation emerges from its physical construction.

Table 5: Voltage outputs of the 'loading' and 'unloading' stages

Input Force[N]	Loading[V]	unloading[V]	Residual
0.00	1.02	1.05	0.030
0.01	1.07	1.09	0.022
0.25	1.10	1.12	0.024
0.26	1.12	1.50	0.024
0.44	1.14	1.163	0.017
0.61	1.17	1.18	0.009
0.84	1.19	1.24	0.040
1.012	1.22	1.25	0.031
1.29	1.24	1.27	0.024
1.62	1.27	1.29	0.020
1.85	1.29	1.31	0.019
2.109	1.32	1.33	0.012
2.45	1.35	1.36	0.01
2.73	1.37	1.38	0.01
3.09	1.34	1.41	0.01
3.5	1.41	1.41	0.00

4.5. Zero offset & mechanical hysteresis

The zero offset is the amount of deviation in sensor's output voltage from the exact value at the lowest point of the measurement range. Zero offset can be expressed in signal units (i.e. Volts), or as a percentage of the full scale range. As it is shown in Equation (1) from the linear regression, the zero offset value for the designed sensor is 1.1045[V]. since the input range is from 0 to 4[N] and output range is 1.1045 [V] to 1.6[V], the zero offset tolerance of the sensor is 1.1045[V] ± 0.1238[V].

Mechanical hysteresis can affect the sensor performance, due to the elastic-plastic deflection of the cantilever beam affecting not only the instantaneous application of the load, but also the previous

application or history. Table 5 shows the voltage outputs during 'loading' and 'unloading'.

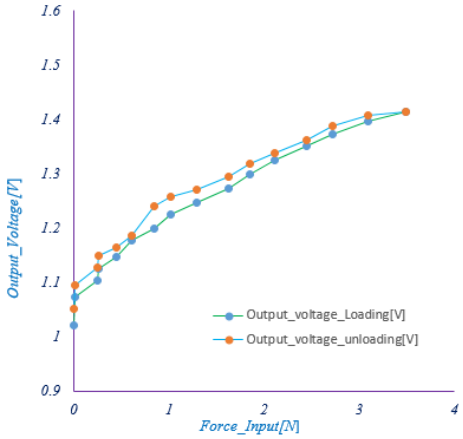


Figure 8: Sensor hysteresis curve

The data of Table 5 were collected during the calibration processes. These values are also plotted within Figure 8. According to these output, the 'loading' and ''unloading' curves provide different values of the sensor output values vs. same value of the applied force (i.e. the input). Therefore, the percentage of the hysteresis van be calculated as the ratio of the maximum difference between the output values curve and sensor range. The effective percentage of the maximum hysteresis error is 9.24%, namely each sensor measurements during unloading will be 9.24% greater from each loading measurements [11]. A summary of the overall sensor performance parameters is reported in Table 6.

Table 6: Summary of the performance parameters of the sensor

Sensor parameter	Value	Unit
Maximum percentage relative error	±0.03109	Full scale output(FS)
Span	[0 - 4]	N
Output range	[1.1045 - 1.6]	V
Resolution	0.01	N
Zero offset	1.1045	V
Zero offset tolerance	±0.123	V
Sensitivity	0.0978	V/N
Maximum Hysteresis error	±9.24%	FS

5. Conclusion and Future work

This works presented simplified design and integration of an optical force sensor to measure external applied force to the biomimetic fingertip during physical interaction. The FEA analysis of the phalange has been conducted to ensure that how much amount of force the fingertip can resist without undergoing plastic deformation. Then the actual finger has been fabricated using HP 3D printing machine. The structural components of the sensor (i.e. cantilever beam, reflective surface and the two optical fibre) were integrated at the fingertip. Sensor calibration and linear regression analysis has been conducted. The regression analysis shows high correlation between applied force and corresponding voltage output from the digital fibre optic sensor. The sensors performance parameters have been briefly addressed. Various applications like surgical operation, manipulations in area of high electromagnetic noise can be benefited from the proposed sensor's performance (i.e. sensitivity, accuracy, linearity and resolution) and its EMC design. As a future work the structural components and arrangement of the sensor can be further optimized to get a better performance and improved sensor characteristics. The deployment of optical fibre, shape of the reflective surface and structure of built-in cantilever beam can be further modified and improved.

Acknowledgment

This paper is an extension of work originally presented in at IEEE Sensors 2017 [12].

References

[1] Konstantinova, J, Stilli, A, Althoefer, K, Force and proximity fingertip sensor to enhance grasping perception. Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on. IEEE, 2015, 2118–2123.
[2] Noh, Y, Sareh, S, Back, J, Würdemann, HA, Ranzani, T, Secco, EL, Faragasso, A, Liu, H, Althoefer, K. A three-axial body force sensor for flexible manipulators. Robotics and Automation (ICRA), 2014 IEEE International Conference on. IEEE, 2014, 6388–6393.
[3] Noh, Y, Secco, EL, Sareh, S, Würdemann, HA, Faragasso, A, Back, J, Liu, H, Sklar, E, Althoefer, K, A continuum body force sensor designed for flexible surgical robotics devices. Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. IEEE, 2014, 3711–3714.
[4] Xu, Z, Todorov, E, Design of a highly biomimetic anthropomorphic robotic hand towards artificial limb regeneration. Robotics and Automation (ICRA), 2016 IEEE International Conference on. IEEE, 2016, 3485–3492.
[5] Salinas, R, 3D Printing with RepRap Cookbook; Packt Publishing Ltd, 2014.
[6] Akin, JE, Finite element analysis concepts: via SolidWorks; World Scientific Publishing Co Inc, 2010.
[7] Xu, Z, Kumar, V, Matsuoka, Y, Todorov, E, Design of an anthropomorphic robotic finger system with biomimetic artificial joints. Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on. IEEE, 2012, 568–574.
[8] Callister, WD, Rethwisch, D.G. Materials science and engineering; Vol. 5, John Wiley & Sons NY, 2011.
[9] Radwin, RG, Oh, S, Jensen, TR, Webster, JG, External finger forces in submaximal five-finger static pinch prehension. Ergonomics 1992, 35, 275–288.
[10] Draper, NR, Smith, H, Pownell, E, Applied regression analysis; Vol. 3, Wiley New York, 1966.
[11] Fraden, J, Handbook of modern sensors; Vol. 3, Springer, 2010.
[12] Agidew TF, Secco EL, Maereg AT, Reid D, Nagar AK, Fibre Optic-Based Force Sensor for Bio-Mimetic Robotic Finger, IEEE Sensors 2017, Glasgow, UK.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 23-32 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Low-Dimensional Spaces for Relating Sensor Signals with Internal Data Structure in a Propulsion System

Catherine Cheung[*,1], Nicolle Kilfoyle[2], Julio Valdés[3], Srishti Sehgal[1], Richard Salas Chavez[1]

[1]*National Research Council Canada, Aerospace, Ottawa, Canada*

[2]*Department of National Defence, Ottawa, Canada*

[3]*National Research Council Canada, Digital Technologies, Ottawa, Canada*

ARTICLEINFO

Article history:
Received: 17 August, 2018
Accepted: 28 October, 2018
Online: 01 November, 2018

Keywords:
Low-dimensional spaces
Condition indicators
Failure prediction
Intrinsic dimension

ABSTRACT

Advances in technology have enabled the installation of an increasing number of sensors in various mechanical systems allowing for more detailed equipment health monitoring capabilities. It is hoped the sensor data will enable development of predictive tools to prevent system failures. This work describes continued analysis of sensor data surrounding a seizure of a turbocharger within a propulsion system. The objective of the analysis was to characterize and distinguish healthy and failed states of the turbocharger. The analysis approach included mapping of multi-dimensional sensor data to a low-dimensional space using various linear and nonlinear techniques in order to highlight and visualize the underlying structure of the information. To provide some physical insight into the structure of the low-dimensional representation, the transformation plots were analyzed from the perspective of several engine signals. By overlaying operating ranges of individual sensor signals, certain regions of the mappings could be associated with distinct operational states of the engine, and several anomalies could be related to various points in the turbocharger seizure. Although the failed points did not map to an obvious outlier location in the transformations, incorporating expert domain knowledge with the data mining tools significantly enhanced the insight derived from the sensor data.

1. Introduction

Rapid developments in sensor technology, data processing tools and data storage capability have helped fuel an increased appetite for equipment health monitoring in mechanical systems. As a result, the number of sensors and amount of data collected for health monitoring has grown tremendously. It is hoped that by collecting large quantities of operational data, predictive tools can be developed that will provide operational, maintenance and safety benefits. Data mining and machine learning techniques are important tools in addressing the ensuing challenge of extracting useful results from the data collected. However, incorporating as much physical domain knowledge to the analysis as possible is also necessary to ensure the results are relevant and practical for the operator and end-user.

This work describes continued analysis of sensor data for the turbocharger subsystem of a diesel engine system. The engine has hundreds of sensors monitoring both the inputs of the engine operator and the resulting equipment outputs. A turbocharger seizure was recorded by the diesel engine sensor system. Therefore, data analysis of this incident including the lead up to the event allows for monitoring and identification of changes in equipment condition indicators with a known outcome.

This paper is an extension of work originally presented at the 2017 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS) [1]. The initial data analysis of this event included intrinsic dimension analysis and relied on clustering techniques for data reduction to transform the high-dimensional sensor data to a low-dimensional space. In this extended version of the paper, the data has not been reduced using clustering techniques in order to minimize information loss and t-Distributed Stochastic Neighbour Embedding (t-SNE) is included as an additional mapping method. An important addition in this paper is further analysis to relate individual sensor signals to the internal data structure of the low-dimensional mappings.

[*]Catherine Cheung, 1200 Montreal Rd, Bldg M-14, Ottawa, ON, Canada K1A 0R6, +1-613-998-1541, Email: catherine.cheung@nrc-cnrc.gc.ca

Failure detection in mechanical systems using data-driven models is an area that has been the focus of much published research in the last decade or so. Machinery failures are hard to predict due to the complex nature of the structure and functions of the system. Using data driven models helps reduce maintenance costs, improve productivity and increase machine availability [2]. For example, fuzzy support vector machines have been applied to identify faults in induction motors [3]. In another case, various data-driven models such as support vector machines, decision trees and kernel methods were used and compared to diagnose faults in shaft and bearings of rotating machinery [4]. Additionally, a model was developed to continuously monitor the health of wind turbine gearboxes [5]. In brief, data-driven models demonstrate great potential for failure detection and preventive maintenance in mechanical applications.

Other related work revolves around understanding the trends data-driven models produce. In several fields, data is being acquired at an astounding rate [6]. There is a significant need for the development of methods and techniques to obtain useful knowledge from these growing sets of available data. Currently, there exist few processes that combine expert knowledge of a system with algorithm-generated prediction models of the system's datasets. However, in cases where expert knowledge is combined with machine learning techniques, there is a notable improvement in the results. For example, a methodology was proposed to detect web attacks [7]. In this procedure, features that represented the tendencies of the system were chosen by experts and were combined with output provided by *n-grams*, a feature extraction algorithm. In another instance, similar to mechanical applications, a framework to combine the clinical intuitions and experience of medical experts with machine learning models was used to overcome the lack of ideal training sets [8]. Medically trained experts provided a task to accomplish, the desired outcome, the data, and helped construct relationships between variables with the algorithm designers. The relationships that were provided aligned with the intuition of the medical expert and their understanding of how factors played out in predicting a certain outcome [8]. However, when expert-knowledge is not always readily available, or if all the variables cannot be identified for a particular outcome, or when proper variable relationships cannot be constructed for a particular outcome, the need to be able to obtain useful knowledge from the results of data-driven models still exists.

In this work, a multi-disciplinary approach to gain knowledge from high-dimensional and voluminous datasets generated by complex real-world systems is explored. Data mining and machine learning techniques are implemented to gather useful insights from the large amounts of sensor data collected for this diesel engine system. By incorporating expert domain knowledge with the low-dimensional representations of the data, a more practical understanding of the data structure presented in the mappings is provided which helps ensure that the results are relevant and accessible to the operator and end-user.

This paper is organized as follows: description of the turbocharger sensor data and the turbocharger seizure are provided in Section 2; details of the implemented data analysis tools and techniques are given in Section 3; the data pre-processing steps and the experimental settings are outlined in Section 4; the intrinsic dimension analysis and low-dimensional mappings are included in Section 5; Section 6 provides the results of several sensor signals overlaid on the visualizations; and finally, Section 7 summarizes the findings of the paper.

2. Turbocharger data

The analyzed turbocharger system contains twin air-cooled turbochargers providing the air-charge to the medium-speed diesel engine system. The diesel engine system consists of two banks of 10-cylinders, denoted Bank A and Bank B. A single turbocharger is assigned to each 10-cylinder bank; the two turbochargers are differentiated as Turbo A and Turbo B, where the letter 'A' or 'B' identifies their respective cylinder bank.

The incident recorded by the engines sensor system and analyzed within this work relates to the seizure of Turbo A [1]. The sensor data recorded for this particular incident was available for the month leading up to and including the time of the incident. From the resulting investigation of the Turbo A seizure, a series of key events leading to the incident were noted. The chronology of the incident's events is detailed below, with the time of occurrence indicated as (hh:mm).

- Noted loss of sensor reading on Turbo A speed sensor
- Engine shut down for inspection
- Turbo A and B speed sensors switched
- Engine restarted at idle, still no Turbo A speed reading indicated (01:12)
- Engaged diesel engine (01:41)
- Higher speed setting requested (01:42), engine exhaust temperatures increased beyond alarm threshold, with no speed increase achieved (01:43 - 01:44)
- Diesel engine disengaged and shut down (01:44 - 01:45)

Following the incident, an inspection of Turbo A was conducted. From the inspection it was determined that the seizure of Turbo A occurred due to a sensor installation error, which occurred when the speed sensors for Turbo A and B were switched. Insufficient spacing between the speed sensor and the turbine's thrust collar led to rubbing and eventually the turbocharger seizure [1].

Although this failure was caused by installation error rather than gradual deterioration of a system element, the ability to characterize and distinguish the healthy state from the seized state of the turbocharger system using data analysis tools is of significant value. This type of analysis could aid in establishing failure models for further predictive work.

2.1. Turbocharger sensor signals

The sensor system of the diesel engine is comprised of 238 sensors that capture information related to operator inputs, equipment outputs, and sensor data. The sensor system provides a means for staff to control system components, monitor the systems status, or be notified via alarm when various pieces of equipment operate outside of pre-set threshold values. In addition, the sensor system allows for the recording and archiving of the operational data measured from various instruments at rates up to 2 Hz. From the 238 sensors relevant to the diesel engine, a subset of 31 signals relating to the operation of the turbochargers was selected. The 31

sensors considered within this analysis, listed in Table 1 [1], encompass parameters such as speeds, temperatures (inlet, outlet, and exhaust), pressures, and shaft torque.

Table 1: 31 turbocharger input parameters

Signal #	Signal
1	Average cylinder exhaust temperature Bank A
2	Average cylinder exhaust temperature Bank B
3	Turbo A speed
4	Turbo B speed
5	Turbo B inlet temperature
6	Turbo B outlet temperature
7	Turbo A inlet temperature
8	Turbo A outlet temperature
9	Charge air manifold pressure
10	A1 cylinder exhaust gas temperature
11	B1 cylinder exhaust gas temperature
12	A2 cylinder exhaust gas temperature
13	B2 cylinder exhaust gas temperature
14	A3 cylinder exhaust gas temperature
15	B3 cylinder exhaust gas temperature
16	A4 cylinder exhaust gas temperature
17	B4 cylinder exhaust gas temperature
18	A5 cylinder exhaust gas temperature
19	B5 cylinder exhaust gas temperature
20	A6 cylinder exhaust gas temperature
21	B6 cylinder exhaust gas temperature
22	A7 cylinder exhaust gas temperature
23	B7 cylinder exhaust gas temperature
24	A8 cylinder exhaust gas temperature
25	B8 cylinder exhaust gas temperature
26	A9 cylinder exhaust gas temperature
27	B9 cylinder exhaust gas temperature
28	A10 cylinder exhaust gas temperature
29	B10 cylinder exhaust gas temperature
30	Shaft torque position 1
31	Shaft torque position 2

The data from the month leading up to and including the turbocharger seizure was analyzed. With the data down-sampled to 1-minute intervals, there were 9968 data points during the time period. The data points prior to the turbocharger seizure were designated as 'healthy', while the points from midnight of the date of the incident were considered 'failed' points. These failed points include all of the data points after the seizure as they correspond to data related to the seized subsystem. As a result, there were 9875 healthy points and 93 failed points identified.

3. Analytical techniques

An important aspect of this workwas the characterization of the relationship between the healthy and failed states of the turbocharger system, particularly the transition between the two states. The original sensor data is described by a multidimensional time-series composed of the 31 signals. Typical from these kind of data is the presence of noise, irrelevancies and redundancies between the descriptor variables, as in reality the core of the data represents a subspace of lower dimension embedded within the higher dimensional descriptor space. In such situations, transformations to lower dimensional spaces are useful for highlighting and visualizing the underlying structure of the information.

To that end, a suitable transformation, preferably with an intuitive metric [9] would produce a mapping of the original high dimensional objects into a lower dimension one, such that a certain property of the data is preserved by the representation. Desired properties characterizing data structure could be conditional probability distributions around neighbourhoods, similarity relations and others. Under normal circumstances, such transformations imply some information loss or error that should be minimized. If successful, the transformation would generate a new set of features out of the original ones which would preserve the desired property, but in a lower dimension representation space that mitigates the curse of dimensionality.

3.1. Intrinsic dimensionality analysis

When choosing the dimension of the target space, it is important to consider the intrinsic dimensionality of the original information which is typically understood as the minimum number of variables required to account for the observed properties of the data [10-13]. Given a functional measure of information loss, it is the minimum number of dimensions (descriptor variables) required to describe the data that minimizes that measure. This concept could be understood in several ways, which results in different algorithms aiming at producing such an estimation. Some approaches focus on local properties of the data, whereas other techniques emphasize the analysis of global properties of the data. Most complex systems in the real world exhibit nonlinear relations among their state variables, which make linear estimators of intrinsic dimensionality at a global scale less powerful than their nonlinear estimation counterparts. However, some of them are computationally expensive.

From the practical point of view, the smaller the choice of the target dimension with respect to the intrinsic one, the higher the representation error would be. On the other hand, choosing values higher than the chosen dimension introduce unnecessary attributes, redundancies and possibly noise. For visualization purposes, three or two dimensions are forcibly required. In these cases, the value of the intrinsic dimension provides a useful guide to the level of caution required when making inferences based on the visualization space.

In this work, the intrinsic dimension of the turbocharger data is estimated using four nonlinear methods and one linear technique: maximum likelihood estimation (MLE), correlation dimension (CD), geodesic minimum spanning tree (GMST), nearest neighbour estimator, and principal component analysis (PCA).

The maximum likelihood estimator [14] is based on the assumption of a Poisson distribution for the k neighbour points and a constant behavior of the probability density function around a given point. The actual estimate of the dimension is derived from the log-likelihood function.

Correlation dimension [15] is one type of fractal dimension and it is one of the most commonly used techniques for estimating the intrinsic dimension. The idea is to compare objects from the point of view of their pairwise distances, producing a normalized count of those pairs whose distance does not exceed a given threshold (the correlation integral). The estimate is given by the slope of a

log-log linear regression of the correlation integral values vs. the different distance thresholds *r*.

The geodesic minimum spanning tree (GMST) estimator [16] assumes that i) the set of multivariate objects are in a smooth manifold embedded within the higher dimensional space determined by the original descriptor variables and ii) these objects are realizations of a random process from an unknown multivariate probability density distribution. This technique produces an asymptotically consistent estimate of the manifold dimension without requiring the reconstruction of the manifold or the estimation of the multivariate distribution of the objects. The first step is to construct a graph based on *k*-neighbourhood density (or neighbourhood distances) where every object is connected with the others nearby. The second step is to build a minimal spanning tree (MST). The distances along its edges and the overall length are used to estimate parameters of the manifold, like entropy and dimension.

The nearest neighbour estimator [17] presents some similarities with the correlation dimension. It is motivated by the possibility of approximating the unknown probability density of the set of multidimensional objects, by normalizing the relative number of nearest neighbours by the volume of the hypersphere containing the objects. The procedure computes the smallest radius *r* required to cover *k* nearest neighbours via a linear log-log regression of the average minimum radius vs. *k*.

Principal component analysis (PCA) is an unsupervised, classical method that is among others, it is used to estimate intrinsic dimensionality. The estimation is simply constructed by obtaining the number of eigenvalues whose relative contributions to the overall variance exceeds a predetermined threshold (e.g. 0.975). Singular value decomposition techniques or diagonalization of covariance/correlation matrices are the typical approaches used for finding the components, which are linear combination of the original set of features. The former approach was used in this paper following the algorithm described in [18].

3.2. Transformation from high- to low-dimensional space

The specificities of the data determine its intrinsic dimension and, in particular, when the estimates are not too different from three, mappings targeting that number of dimensions could portray appropriate representations of the data. In these cases, the visualizations obtained with different mapping techniques typically exhibit low errors or information loss measures. They would reveal patterns corresponding to valid relationships within the data like regularities, showing up as clustering structure, as well as abnormalities, less frequent elements and outliers.

Clearly, for machine learning purposes, mitigating the curse of dimensionality is important and often the dimension of the target spaces must go beyond the ones required by visual inspection. In this paper, Principal Components, Sammon mapping, t-SNE and Isomap were used as representatives of linear and nonlinear transformation techniques.

Principal Components

A low-dimensional representation of the data can be produced using the first few principal components found through principal component analysis, which are mutually orthogonal (described in Section 3.1). Their main features were presented in the previous section. From the point of view of visually inspecting the data, the first few principal components (up to three) are used as a baseline low-dimensional representation. They are linearly uncorrelated and the amount of variance contained in each new component decreases monotonically. However, the cumulated variance contained in the first three components is not sufficiently high and it is a crucial element to consider when working with principal components visualizations.

Sammon Mapping

The idea of constructing low dimensional spaces where the distance distribution maximally matches the one in the original space is very intuitive and has been at the core of multi-dimensional scaling methods (MDS) [19-22]. Different variants of this approach have been used for creating visual representations of metric and non-metric data. On representations that aim at preserving distances in the original and the target spaces, nearby/distant objects in the original data space are placed at nearby/distant locations from each other in the low-dimensional target space. Some variants preserve the actual distance values, while others aim at preserving their ranks or their ordering relation.

In the first case, measures based on squared differences between dissimilarities on both spaces are commonplace and are variations of objective functions like

$$\sum_{1\le i,j\le N} w_{ij}\left(F\left(\delta_{ij}^{p}\right)-d_{ij}^{p}\right)^{2}$$

where N is the number of objects, w_{ij} is a weight associated to every pair of objects *i, j*, F is a monotonically increasing function, δ_{ij} is a dissimilarity measure between objects *i, j* in the original data space and d_{ij} is their dissimilarity/distance in the target space, with p as an exponent of the difference term.

From this general formulation, several mapping techniques are derived, in particular, Sammon's nonlinear mapping [23], conceived as a transformation of vectors of two spaces of different dimension ($D > m$) by means of a function $\varphi: R^D \to R^m$, which maps vectors $\vec{x} \in R^D$ to vectors $\vec{y} \in R^m$, $\vec{y} = \varphi(\vec{x})$. The actual objective function to minimize is given by Equation 1:

$$Sammon\ error = \frac{1}{\sum_{i<j}\delta_{ij}} \sum_{i<j} \frac{\left(\delta_{ij} - d(\vec{y}_i,\vec{y}_j)\right)^2}{\delta_{ij}}, \quad (1)$$

where typically d is an Euclidean distance in R^m. The weight term δ_{ij}^{-1} highlights the importance of smaller distances and therefore, the behavior around close neighbourhoods exerts a larger influence on the error function.

t-SNE

A probabilistic principle is used by the Stochastic Neighbour Embedding (SNE) [24], where the goal is to preserve neighbour identities. A dissimilarity or distance between the objects in the original space is used for creating an asymmetric probability for each object with respect to its potential neighbours, with a pre-set neighbourhood notion (the perplexity). The same is performed for the objects in the target space.

The goal is to match the two distributions as much as possible, which is achieved by minimizing the sum of Kullback-Leibler divergences. The rationale is to center a Gaussian on each object in the original space and to use the distances (or given similarities) for constructing a local probability density function on the neighbourhood. The same operation is performed in the transformed, low dimensional space and the purpose is to match the two as much as possible, so that neighbourhoods are preserved.

The t-Distributed Stochastic Neighbour Embedding (t-SNE) [24-26] is an improvement of the original SNE. There are two main distinguishing features of t-SNE: *i)* a simpler symmetric objective function is introduced; and *ii)* instead of a Gaussian distribution, a t-Student distribution is used for the points in the low-dimensional space. These modifications allow for better capturing the local structure of the high-dimensional data and also revealing the presence of clusters at several scales, as indicators of global structure.

Isomap

Isomap [27-30] is a flexible technique oriented to learn non-linear manifolds and overcomes some difficulties inherent to classical methods like principal components or MDS-related. In contrast to the latter, the Isomap technique aims to preserve pair-wise geodesic (or curvilinear) distances rather than plain (Euclidean) ones. Geodesic distances are those measured along the low-dimensional manifold containing the data and therefore, not necessarily objects that are close in the Euclidean sense will be so when geodesic distances are considered.

There are three steps in the procedure: i) build a graph (the neighbourhood graph) that connects all points according to their pair-wise Euclidean distances; ii) estimate the geodesic distances between all pairs of points by calculating their shortest path distances in the neighbourhood graph; and iii) compute a geodesic distance preserving mapping using MDS with Euclidean distance as metric for the low-dimensional space.

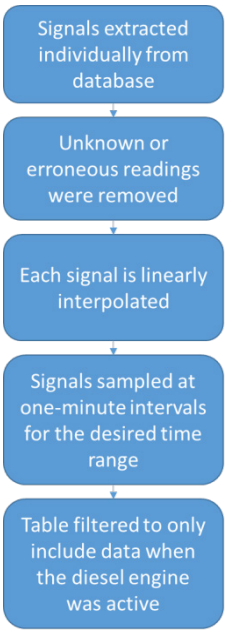
4. Experimental settings

For the investigated time period leading up to and including the turbocharger seizure, there were 9968 data points. These points were sampled at intervals of one-minute. Of the 9968 total points, 9875 were designated as 'healthy' and the remaining 93 designated 'failed'.

To determine the low-dimensional transformation using the Isomap method, 12 nearest neighbours were specified. For the t-SNE transformation, the perplexity was set at 30.

4.1. Data pre-processing

The engine sensor system was originally set up for real-time equipment health monitoring, and not specifically for maintenance or safety purposes. As such, a number of data pre-processing and data consolidation steps were necessary before implementing the data analysis tools. The data pre-processing steps followed are illustrated in Figure 1.



From the full signal database, each of the 31 signals was extracted separately, then unknown and erroneous readings were removed. Afterwards, each signal was linearly interpolated and sampled at one-minute intervals for the desired time range, ensuring that the time range fell within the interpolated values. Finally, a filter was applied to only consider data recorded during active operation of the diesel engine.

The input parameters were standardized in order to fairly compare variables measured in different units and different ranges. The standardized variables had a mean value of zero and standard deviation equal to one.

5. Low-dimensional mappings

5.1. Intrinsic dimension results

Estimates of the intrinsic dimension of the turbocharger data were calculated using the five techniques detailed in Section 3.1. These estimates are listed in Table 2. The first three principal components represent 0.981 of the total variance in the data.

Table 2: Intrinsic dimension estimates

Estimation Method	Estimate
Maximum Likelihood Estimator	5.239
Correlation Dimension	1.285
Geodesic Minimum Spanning Tree	4.202
Nearest Neighbour Dimension	0.307
Principal Component Analysis Eigenvalues	3.000

Almost all of the estimates fall in the range of 1 to 5. Since the original sensor space corresponds to a dimension of 31, these estimates show that the information contained in that 31-D high dimensional space could be sufficiently explained more simply by the combination of a few factors. A target dimension of 3 was selected, taking into account the range of estimates.

5.2. Transformation to 3-dimensions

From the intrinsic dimension results, low-dimensional representations of the turbocharger data with three dimensions

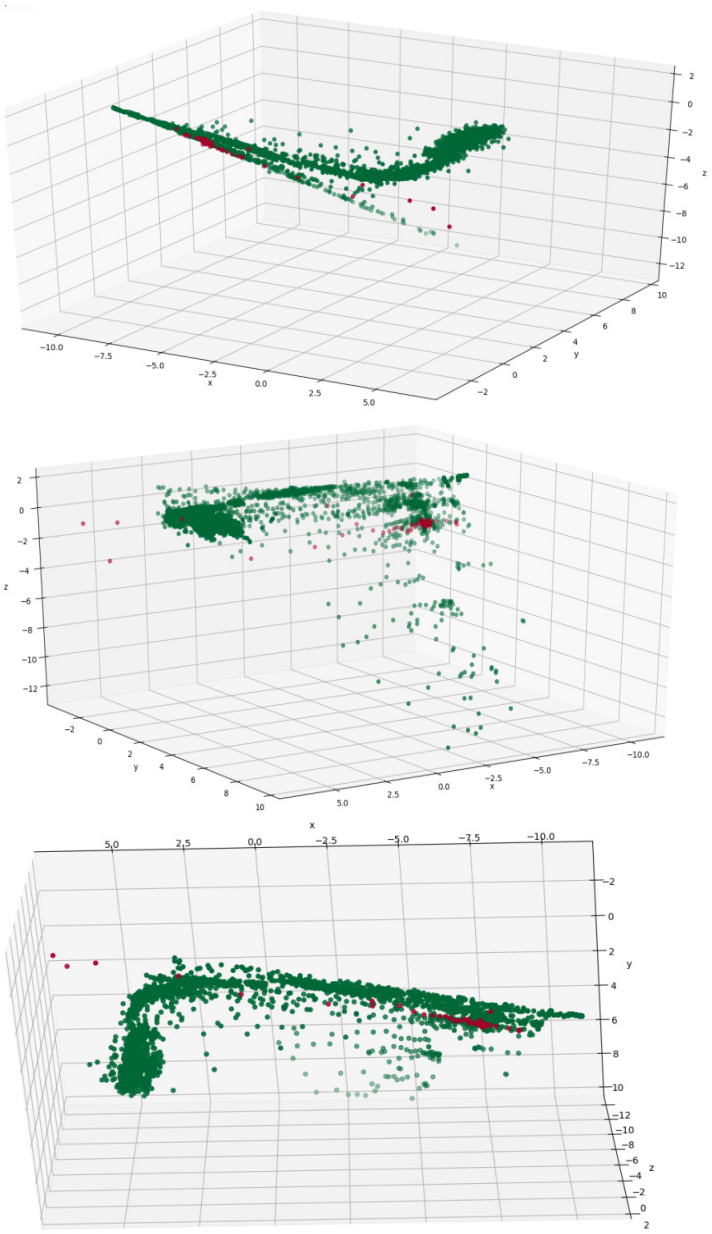


Figure 2: PCA mappings to 3-D

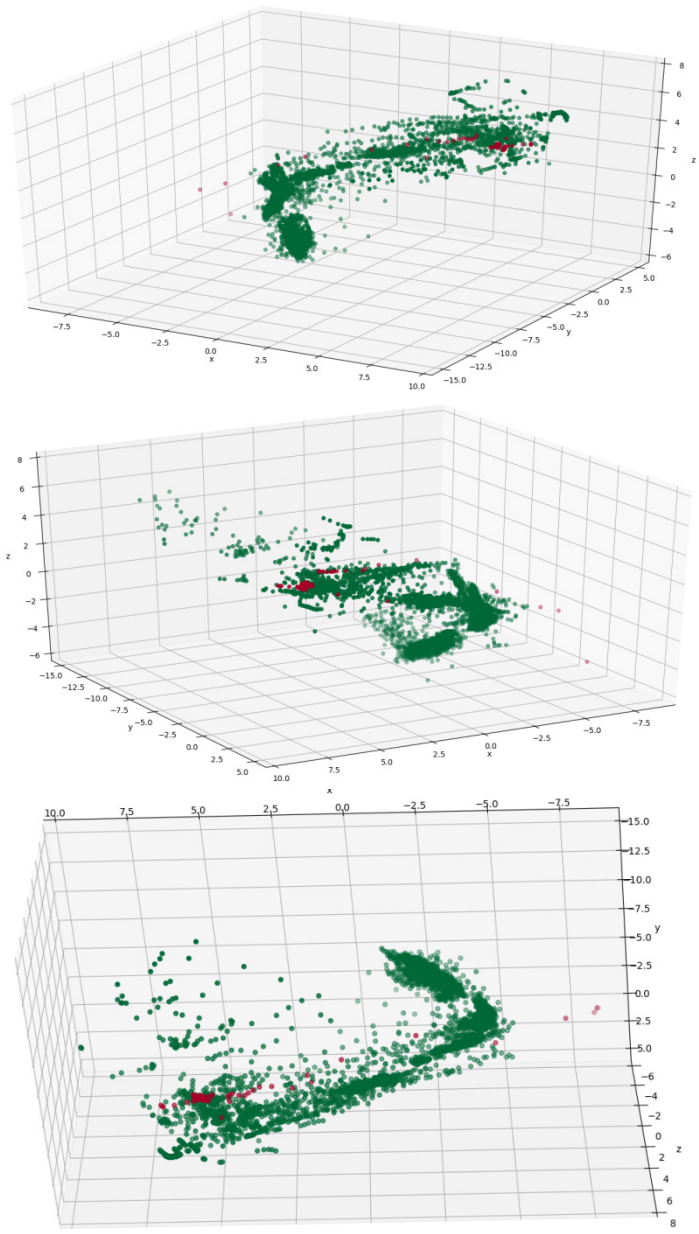


Figure 3: Sammon mappings to 3-D

were sought. The transformation of the original 31-D space, corresponding to the 31 turbocharger sensors, to 3-dimensions was performed using the four methods described in Section 3.2 (principal component analysis, Sammon, t-SNE, and Isomap). The full data set from the time period around the turbocharger event was mapped, consisting of 9968 total points of which 93 were designated 'failed' and the rest 'healthy'.

Images of the 3-D spaces obtained from the different mappings are presented in Figures 2-5. In these images, the healthy points are coloured green, while the failed points are red. Since the ratio of healthy to failed points is tremendously imbalanced (9875:93), it may be difficult to see the failed points. Representing 3-D scenes by 2-D images is clearly not ideal because of the limitations in exploring different perspectives and distances between objects in the scene. Several snapshots of the 3-D space are included to help overcome that limitation.

Figure 2 shows three views of the low-dimensional mapping using the first three principal components determined through PCA. These three components are orthogonal and are linear combinations of the 31 turbocharger variables. Figure 3 depicts three views of the Sammon mapping to 3-dimensional space. Figure 4 shows the t-SNE mapping to 3-D. Figure 5 illustrates the Isomap transformation to 3-dimensions.

In the PCA mapping (Figure 2), the data is distributed mostly in a 2-dimensional plane in a boomerang-like shape. The failed points are concentrated at one end of that boomerang shape ($-5 \leq x \leq 10$). The Sammon mapping (Figure 3) also shows the distribution of the data in a boomerang-like shape. Again, the failed points are clustered at one end of that shape ($3 \leq x \leq 7$).

The t-SNE transformation (Figure 4) has a markedly different data structure with the data organized in more distinct clusters as opposed to the continuous distribution of data in a particular shape

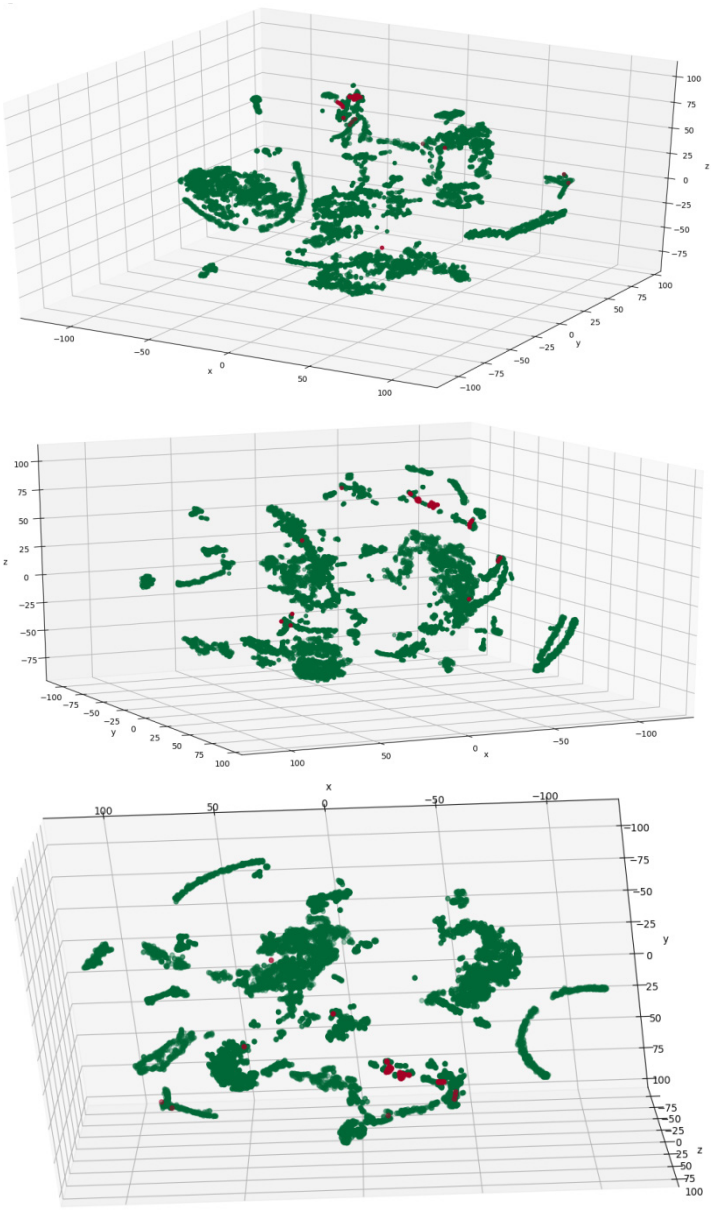


Figure 4: t-SNE mappings to 3-D

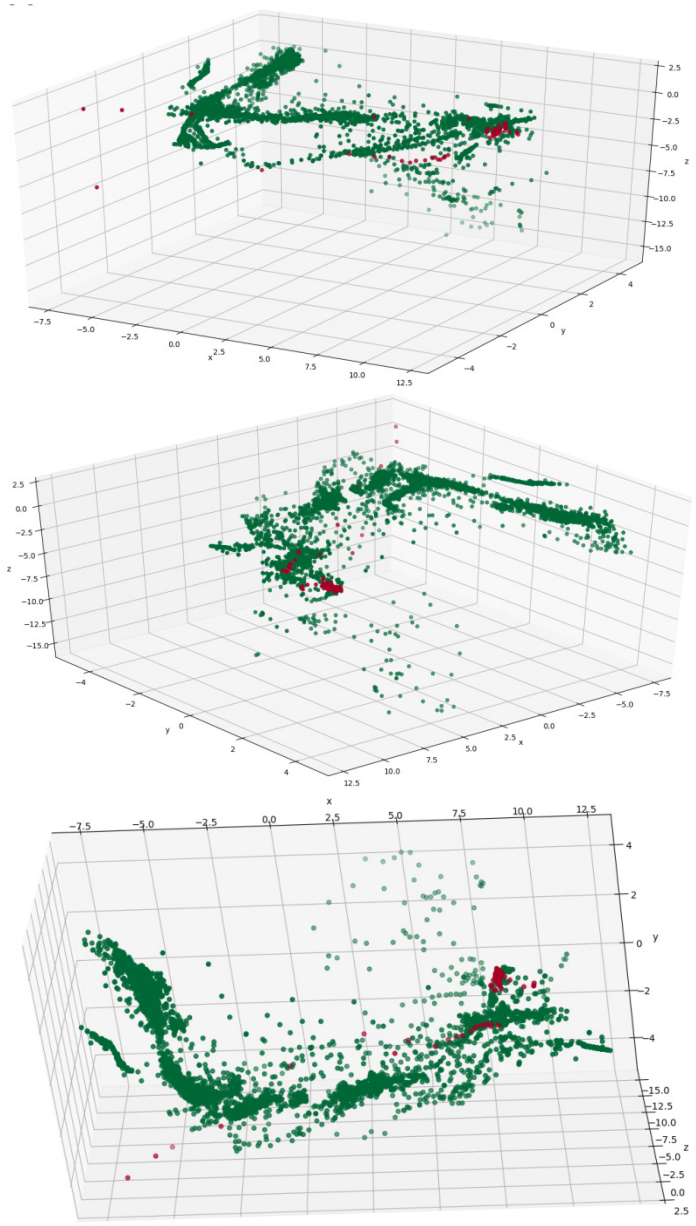


Figure 5: Isomap mappings to 3-D

(e.g. boomerang). The failed points are located in several areas of the structure in the t-SNE mapping. This is a consequence of the property expressed by the mapping (Section 3.2, t-SNE), which focuses on preserving conditional probability distributions within neighbourhoods, rather than distances. Exposing cluster structure more clearly is one of the strengths of t-SNE.

The Isomap plots (Figure 5) show a similar data structure to PCA and Sammon with much of the data falling along a boomerang-like shape. Also similar to PCA and Sammon, the failed points are all concentrated at one end of that main boomerang shape ($5 \leq x \leq 10$). However, more prevalent in the Isomap plot, than with the other plots, is one dense island of healthy data points just offset from that main shape, which will be discussed further in Section 6.2. The similar data structure seen with the PCA, Sammon and Isomap plots can be attributed to the fact that nonlinear effects are mild.

The transformation of the 31-dimension space representing the 31 turbocharger parameters to a low-dimensional 3-D space shows a distinct data structure in the various mapping techniques. The failed points are not mapped to an obvious outlier location in the transformations that can be easily distinguished from the healthy points.

6. Relating sensor signals to internal data structure

In order to better understand these data structure visualizations in a more physical sense, e.g. from the perspective of the operator or maintainer of the engine, the next step was to select a handful of the engine signals and examine how each of these signals impacted and was represented in the data structure. One of the aims was to determine if certain regions of the mappings could be associated with distinct operational states of the engine. Another aim was to demonstrate that individual signals were appropriately represented in the low-dimensional mapping.

6.1. Selected signals

As detailed in Table 1, a variety of sensors recording speeds, temperatures, pressures and torque were included in the analysis. A small subset of signals was selected to investigate the internal data structure of the mappings. The selected signals were: Turbo A speed, Turbo B speed, A1 cylinder exhaust gas temperature, B2 cylinder exhaust gas temperature, and engine speed.

As there was a known failure of the Turbo A speed signal, the Turbo B speed signal was selected to compare the recorded values of the two turbos. In a similar manner, at the time of failure, the exhaust cylinder temperature sensors went into alarm. To enable comparison of the exhaust temperatures, one cylinder exhaust temperature signal was selected from each bank. Finally, engine speed was selected, although it was not part of the 31 turbocharger parameters used for the transformations. This sensor was included as it is the main driving factor for change in the values of the other sensor signals.

6.2. Representation of signals in low-dimensional mappings

Each of the low-dimensional mappings was then overlaid with a heat map of one of these sensor signals. Where instead of the data points coloured green and red for healthy and failed points respectively, they were coloured according to the relative value of the particular signal in its operating range, with blue corresponding to the low end of the range and red for the high end of the operating range.

In the following section of the paper, two regions of interest are highlighted: i) the points surrounding the time frame where the Turbo A speed sensor experienced a fault and failed to record; and ii) the points surrounding the five minutes to the turbocharger seizure.

These two regions were chosen for further investigation because they appear to stand out the most in the visualizations. The Isomap low-dimensional mappings are shown in this paper with an overlay of the heat map for the five selected signals. Figure 6

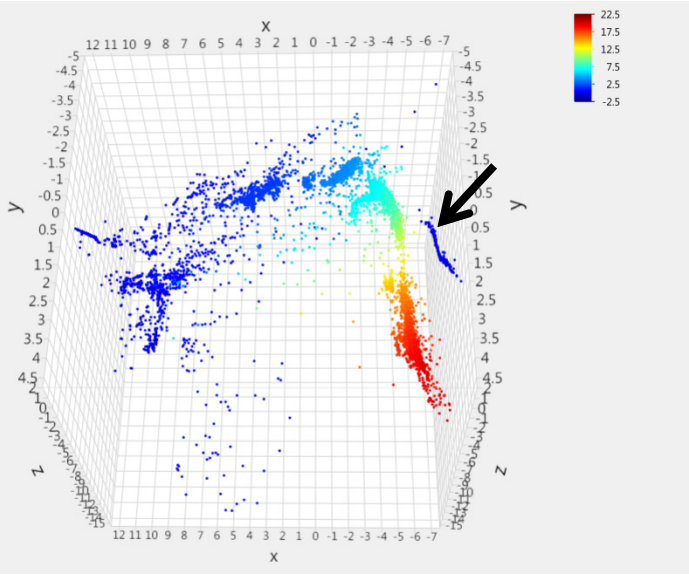


Figure 6: Turbo A speed sensor value [kRPM] overlay onto Isomap transformation

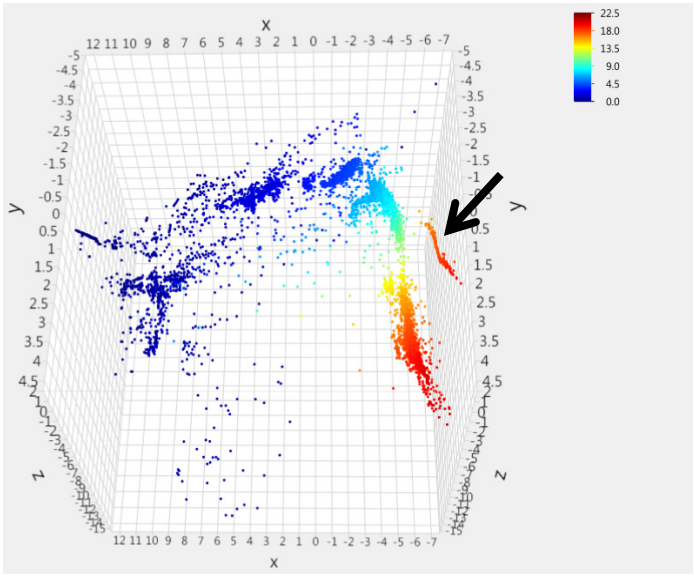


Figure 7: Turbo B speed sensor value [kRPM] overlay onto Isomap transformation

shows the overlay of the Turbo A speed signal on the Isomap transformation. Figure 7 shows the overlay of the Turbo B speed signal on the Isomap transformation. Since both these signals are speed sensors on turbochargers, the expectation is that these two plots should be very similar. For the most part the two plots are indeed very similar, however, it should be noted that the speed ranges are slightly different in the legend.

The distribution of the data points in the mapping indicates that almost all the near-zero/low-speed data points are found in one region of the data structure ($x \geq -2$) while the data points with high speed values are concentrated in a different region ($-7 \leq x \leq -5$). There is a strong gradient between these two regions containing a transition region of the intermediate speeds.

However, an abnormality to the general structure of the data is evident in the high power region of the plots. In Figure 6, the heat map indicates zero values for the Turbo A speed sensor for the data points in the area indicated by the arrow. In the same area, indicated by the arrow in Figure 7, the Turbo B speed sensor portrays high speed values. The overlays of the other examined signals displayed expected values for those data points in that high power region. Thus this group of points is likely related to the loss of the Turbo A speed sensor. Indeed the data points in this area were later found to coincide with the known initial loss of the Turbo A speed sensor readings.

Figure 8 shows the overlay of the Turbo A Cylinder 1 exhaust gas temperature signal on the Isomap transformation. Figure 9 shows the overlay of the Turbo B Cylinder 2 exhaust gas temperature signal on the Isomap transformation. Figure 10 shows the overlay of engine speed signal on the Isomap transformation. The plots of the two cylinder exhaust temperatures from each bank (Figures 8 and 9) are quite similar to each other, as one would expect. They are also structured similar to the Turbo speed overlays where the data points with low temperature values also have low speed values, while the high temperature points have high speed values.

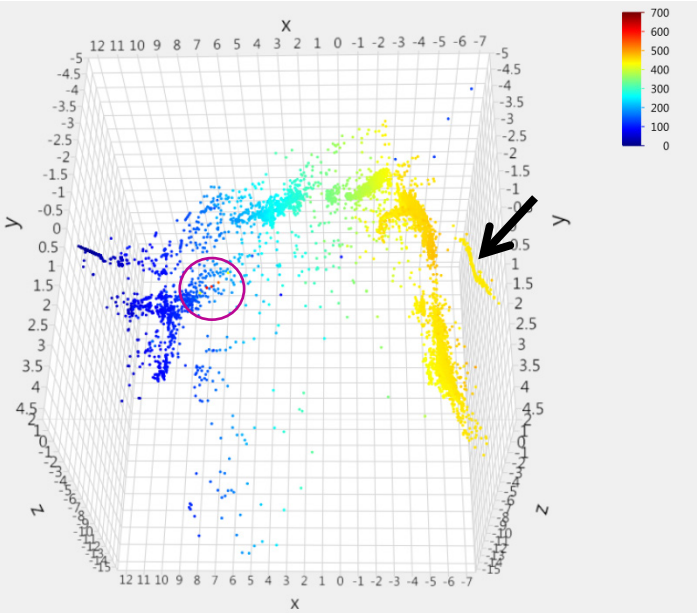


Figure 8: Turbo A Cylinder 1 exhaust gas temperature [˚C] overlay onto Isomap transformation

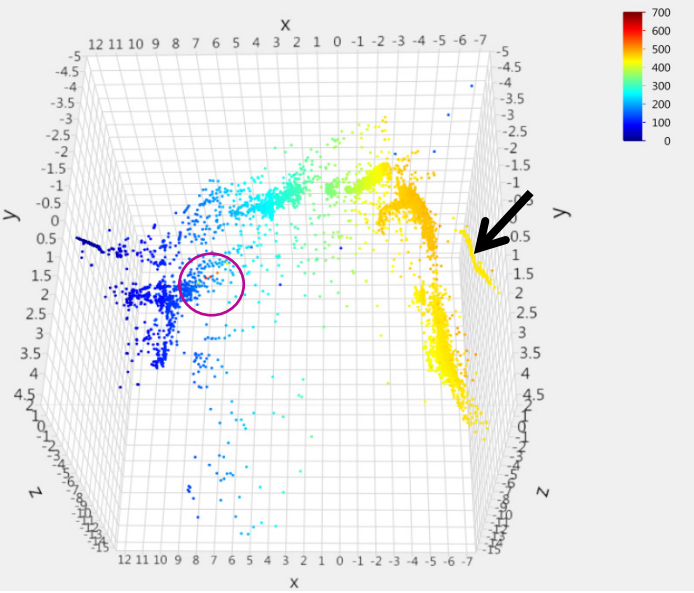


Figure 9: Turbo B Cylinder 2 exhaust gas temperature [˚C] overlay onto Isomap transformation

A small cluster of high cylinder temperature indications is found within the low operating range group indicated by the purple circle in Figures 8 and 9. These points were later found to be the points occurring five minutes before the turbocharger seizure. The location of these points in the low speed range is consistent with the findings of the turbocharger incident report (described in Section 2), where an increase in speed was requested but no increase in speed was realized although the exhaust gas temperatures rose to high levels invoking the alarms.

Although engine speed was not a signal used to generate the Isomap transformation, here it is used as a means to support the previously observed trends of low and high operating ranges in the plots. From Figure 10, the engine speed overlay demonstrates a similar distribution of the low and high engine speeds in the data

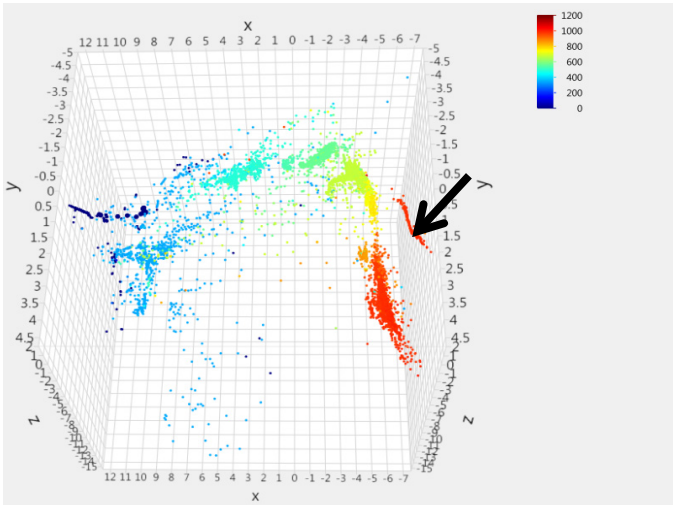


Figure 10: Engine speed [RPM] overlay onto Isomap transformation

structure. Looking again at the data points indicated by the arrow in Figure 10, the trend seen previously with the Turbo speeds and cylinder exhaust temperatures is confirmed.

According to Figure 5, the data points indicated by the arrow in the prior figures are clearly distant from the main set of data and are still designated as 'healthy'. It was also pointed out that the small cluster of cylinder exhaust gas temperature indications (some of the points inside the region of the purple circle in Figures 8 and 9) represent values that were recorded between five hours and five minutes to the turbocharger seizure. These points embed themselves inside low operating ranges of these signals and are also designated as 'healthy' data (Figure 5). In both cases, the aforementioned groups of points display interesting traits through the unique combination of their healthy/failed designation, the value of the overlaid signal that they show, and their distance from other groups of data points. This demonstrates that the failure of the turbocharger system may not always be associated with extreme signal values, as one may assume. These traits help uncover new trends in the data structure, like the abnormality indicated by the arrow or the cluster of points within the purple circle in the preceding figures, which should be investigated. Thus, being able to combine expert domain knowledge of sensors and failures with data mining tools is an invaluable method of extracting and understanding information from sensor measurements.

7. Concluding Remarks

This work describes continued analysis of sensor data for the turbocharger subsystem of a diesel engine system. The engine has hundreds of sensors monitoring both the inputs of the engine operator and the resulting equipment outputs. The objective of the data analysis was to characterize and distinguish the healthy and failed states of the turbocharger seizure as recorded by the diesel engine sensor system. The analysis approach included the mapping of high-dimensional sensor data to a low-dimensional space using a variety of linear and nonlinear techniques in order to highlight and visualize the underlying structure of the information.

Estimates of the intrinsic dimension were obtained to determine the appropriate number of dimensions required by the

low-dimensional transformations and to guide the interpretation of the visualization spaces produced. For this case, three dimensions was an appropriate estimate. Through the unsupervised process of these transformations, the structure of the turbocharger data could be visualized and inspected. The transformation methods included principal components, Sammon mapping, t-Distributed Stochastic Neighbour Embedding, and Isomap. The transformation of the 31-dimension space representing the 31 turbocharger parameters to a low-dimensional 3-D space shows a distinct data structure in the various mapping techniques. The failed points are not mapped to an obvious outlier location in the transformations that can be easily distinguished from the healthy points.

In order to gain more physical insight into the internal data structure of the resulting mappings, the transformation plots were analyzed from the perspective of several engine sensor signals. By overlaying operating ranges of individual sensor signals, certain regions of the mappings could be associated with distinct operational states of the engine. Low and high operating engine regions could be clearly seen in the internal data structure, and several anomalies could be identified which were then associated to various points in the turbocharger seizure. These results are extremely promising and demonstrate how operational knowledge can be easily incorporated with the data analytics tools to enhance the insights that can be gained from the sensor measurements.

In this work, data mining and machine learning techniques are implemented to gather useful insights from the large amounts of sensor data collected for this diesel engine system. By incorporating expert domain knowledge with the low-dimensional representations of the data, a more practical understanding of the data structure presented in the mappings is provided which helps ensure that the results are relevant and accessible to the operator and end-user.

Future efforts are aimed at expanding this analysis to data from other diesel engines and other failures in the engine system. Further work to generalize the analysis to a diesel engine system model instead of a turbocharger-specific model is in progress. Efforts are also underway to better characterize the healthy and failed states, through classification and anomaly detection techniques. The development and implementation of these tools should help enable advance indication of a change in behavior that could be investigated before a major incident.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge and thank Defence Research and Development Canada for their support of this research.

References

[1] C. Cheung, J. J. Valdés, A. Lehman Rubio, R. Salas Chavez, C. Bayley, “Low-dimensional spaces for the analysis of sensor network data: identifying behavioural changes in a propulsion system" in 5th International Symposium on Robotics and Intelligent Sensors, Ottawa, Canada, 2017. https://www.doi.org/10.1109/IRIS.2017.8250133

[2] A. Widodo, B. Yang, "Support vector machine in machine condition monitoring and fault diagnosis" Mech. Syste. Signal Pr., 21(6), 2560-2574, 2007. https://www.doi.org/10.1016/j.ymssp.2006.12.007

[3] S. Li, "Induction motor fault diagnosis based on fuzzy support vector machine" in 3rd International Conference on Electromechanical Control Technology and Transportation, 2018. https://www.doi.org/10.5220/0006975105880592

[4] M. Saimurugan, K. Ramachandran, V. Sugumaran, N. Sakthivel, "Multi component fault diagnosis of rotational mechanical system based on decision tree and support vector machine" Expert Syst. Appl., 38(4), 3819-3826, 2011. https://www.doi.org/10.1016/j.eswa.2010.09.042

[5] M.C. Garcia, M.A. Sanz-Bobi, J.D. Pico, "SIMAP: intelligent system for predictive maintenance" Comput. Ind., 57(6), 552-568, 2006. https://www.doi.org/10.1016/j.compind.2006.02.011

[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, "From data mining to knowledge discovery in databases" AI Mag., 17(3), 1996. https://doi.org/10.1609/aimag.v17i3.1230

[7] C. Torrano-Gimenez, H.T. Nguyen, G. Alvarez, K. Franke, "Combining expert knowledge with automatic feature extraction for reliable web attack detection" Secur. Commun. Netw., 8, 2750–2767, 2015. https://www.doi.org/10.1002/sec.603

[8] F. Kuusisto, I. Dutra, M. Elezaby, E.A. Mendonça, J. Shavlik, E.S. Burnside, "Leveraging expert knowledge to improve machine-learned decision support systems" AMIA Jt Summits Transl Sci Proc, 87–91, 2015.

[9] J. Valdés, C. Y. S. Létourneau, “Data fusion via nonlinear space transformations" in 1st International Conference on Sensor Networks and Applications, San Francisco, USA, 2009.

[10] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press Professional Inc., 1990.

[11] E. Facco, M. d’Errico, A. Rodriguez, A. Laio, "Estimating the intrinsic dimension of datasets by a minimal neighborhood information" Sci. Rep.-UK, 7(1), 2017. https://www.doi.org/10.1038/s41598-017-11873-y

[12] D. Granata, V. Carnevale, "Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets" Sci. Rep.-UK, 6(1), 2016. https://doi.org/10.1038/srep31377

[13] P. Campadelli, E. Casiraghi, C. Ceruti, A. Rozza, "Intrinsic dimension estimation: Relevant techniques and a benchmark framework" Math. Probl. Eng., 2015. https://doi.org/10.1155/2015/759567

[14] E. Levina, P. J. Bickel, “Maximum likelihood estimation of intrinsic dimension" in 17th Conference in Neural Information Processing Systems, Vancouver, Canada, 2004.

[15] P. Grassberger, I. Procaccia, “Measuring the strangeness of strange attractors" Physica D, 9(1-2), 189-208, 1983. https://www.doi.org/10.1016/0167-2789(83)90298-1

[16] J.A. Costa, A.O. Hero, “Geodesic entropic graphs for dimension and entropy estimation in manifold learning" IEEE T. Signal Proces., 52(8), 2210–2221, 2004. https://www.doi.org/10.1109/tsp.2004.831130

[17] K. Pettis, T. Bailey, A. Jain, R. Dubes, “An intrinsic dimensionality estimator from near-neighbor information" IEEE T. Pattern Anal., 1(1), 25-37, 1979. https://www.doi.org/10.1109/tpami.1979.4766873

[18] W. Press, S. Teukolsky, W. Vetterling, B. Flannery, Numerical Recipes in C, Cambridge University Press, 1992.

[19] J. Kruskal, “Nonmetric multidimensional scaling: a numerical method" Psychometrika, 29(2), 115-129, 1964. https://www.doi.org/10.1007/bf02289694

[20] J. Kruskal, “Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis" Psychometrika, 29(1), 1-27, 1964. https://www.doi.org/10.1007/bf02289565

[21] I. Borg, P. Groenen, Modern multidimensional scaling - theory and applications, Springer Series in Statistics, 1997.

[22] I. Borg, P.J.F. Groenen, P. Mair, Applied Multidimensional Scaling, Springer Verlag, 2013.

[23] J. W. Sammon, “A nonlinear mapping for data structure analysis" IEEE T. Comput., 18(5), 401-409, 1969. https://www.doi.org/10.1109/tc.1972.5008933

[24] G. Hinton, S. Roweis, “Stochastic neighbor embedding" Adv. Neur. Inf. Proc. Systems, 15, 857-864, 2003.

[25] L. van der Maaten, G. Hinton, “Visualizing data using t-sne" J. Mach. Learn. Res., 9, 2579-2605, 2008.

[26] M. Nguyen, S. Purushotham, H. To, C. Shahabi, "M-tsne: a framework for visualizing high-dimensional multivariate time series", University of Southern California, 2017. https://arxiv.org/abs/1708.07942

[27] J. Tenenbaum, V. de Silva, J. Langford, “A global geometric framework for nonlinear dimensionality reduction" Science, 290(5500), 2319-2323, 2000. https://www.doi.org/10.1126/science.290.5500.2319

[28] M. Bernstein, V. de Silva, J. Langford, J. Tenenbaum, “Graph approximations to geodesics on embedded manifolds" Stanford University, Tech. Rep., 2000.

[29] V. de Silva, J. Tenenbaum, “Global versus local methods in nonlinear dimensionality reduction" Adv. Neur. Inf. Proc. Systems, 15, 721-728, 2003.

[30] H. Choi, S. Choi, "Robust Kernel Isomap" Pattern Recogn., 40(3), 853-862, 2007. https://www.doi.org/10.1016/j.patcog.2006.04.025

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 26-32 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Stream Cipher by Reed-Solomon Codes

Tao Wu*[1,2], Ruomei Wang[1]

[1] *School of Data and Computer Science, Sun Yat-Sen University, Guangzhou 510006, P.R. China*

[2] *Shenzhen Research Institute of Sun Yat-Sen University, Shenzhen 518057, P.R. China*

ARTICLEINFO

Article history:
Received: 04 May, 2018
Accepted: 07 July, 2018
Online: 14 July, 2018

Keywords:
Stream cipher
Reed-Solomon code
Public-key cryptography

ABSTRACT

Stream cipher can be used in constrained environments to provide information security and reduce energy expense at data transmission. In this paper, it is shown that Reed-Solomon (RS) code can be used to implement stream cipher, which is widely used for error corrections of data in transmission and storages. The proposed stream cipher combine the keys with the messages, and conceals the cipher within the erroneous RS code. Then at the receiving part the cipher can be checked out by RS decoding, leaving the information unchanged. There are several advantages with this scheme: First, compared with the usual stream cipher including two engines, only one engine is required, making the synchronization of the stream cipher and messages easily. In the situation, the stream cipher and the messages are held as a whole while the alignment of stream cipher is not needed. Thus the usual power consumptions for synchronization and key generation at the receiving part can be saved. Second, the stream cipher increases the security level by adding the difficulty of decoding a random linear RS code, which stores the key streams like public key cryptography. Since the RS encoding and decoding modules are already included in many computer systems or devices, the new scheme may be implemented by reconfiguration rather than extra hardware units. Compared with generalized RS code for code-based public-key cryptography like McEliece system, it uses systematic encoding instead of nonsystematic encoding, which decreases the power overhead. Unlike the encryption by public keys and decryption by private keys of McEliece system, it encrypts messages by private keys from stream cipher and decrypts texts by public keys implemented in the RS decoder.

1 Introduction

Stream cipher by Reed-Solomon (RS) code this work intends to implement encryption by encoding if necessary [1]. As the encoding-encryption paradigm has been used in the standard for mobile telephony GSM [2, 3], it has received much attention so far. RS code has been widely used in communications and storage systems to correct errors, which can also be used together with BCH code to ensure data integrity [4]. The idea of using RS code as a carriage for stream cipher comes from the observation of RS code for code-based cryptography, where it is found that the errors are intentionally inserted and recovered with much computation and energy [5, 6]. In fact, the code-based peudo random generator and its security can be considered to achieve stream cipher [7]. In [8], the generalized RS code is used to reduces the density of a transform matrix and improves weakly secure data exchange.

Stream cipher belongs to symmetric cryptography and can be built by many devices, such as linear and nonlinear feedback shift registers. For example, the Rakaposhi stream cipher consists of a 128-bit nonlinear feedback shift register, a 192-bit dynamic linear feedback shift register, and a nonlinear filter function [9]. Ubiquitous environments demands security, speed and power consciousness in processing huge amount of multimedia data [10], in which stream cipher can be used. Stream cipher is an engine that generates bit streams to mix the message. At the sending part, the

*Tao Wu, Shenzhen Institute of Industry-University-Research, Sun Yat-Sen University, email: nstrch@outlook.com

https://dx.doi.org/10.25046/aj030404

plain text is confused with one bit stream, while at the receiving part, the cipher text is processed with the same bit stream again to recover the message. The stream cipher is like a determined peudo-random discrete function, of which the security can be analyzed by the generating mechanism of the cipher. A stream cipher can be marked by a key stream (K) and the initialization vector (IV), while the adversary may recover all the secret bits of K after observing many related (K, IV) pairs [9].

Usually, an one-way function and multiple steps are devised to safeguard the key generations or initialization processes, let bit XOR as the only operation to encrypt the message. By contrast, people also seek approaches to implement more secure code by public key cryptography. Code based cryptography [11] with Goppa codes [12] or MDPC codes [13], for example, can be applied to realize McEliece system . Meanwhile, generalized RS code can also be used to implement asymmetric cryptography [6], which brings out short keys compared with that obtained by Goppa code.

In this work, a stream cipher with RS code is proposed, in which the messages are encoded as RS codewords. This idea comes from its similarity to the configured logic with a DSP unit in FPGA. At the receiving part, the plaintexts and cipher are recovered from RS codes. The contributions of this paper include:

- a mechanism to include stream cipher in RS encoding and decoding, which is able to cover the message in a simple way.
- a review of related work about stream cipher of different schemes.
- a hardware implementation of stream cipher along by RS code.

The remaining parts of this paper is organized as follows: Sect. 2 is a review of RS encoding and decoding; Sect. 3 briefly introduces the related work with stream cipher and code-based cryptography; Sect. 4 discusses the stream cipher along with RS code; then the hardware implementation and comparison of the stream cipher are given in Sect. 5; the last section concludes this paper.

2 RS Code

RS code is widely used for error corrections. The encoding can be either implemented by polynomial division with shift registers, or it can be performed by multiplying the generating matrix [14]. Suppose the RS code is defined over $GF(2^m)$ with n code bits and k information bits. Message sequence $(u_{k-1}, \ldots, u_1, u_0)$ is equivalent to

$$u(X) = u_0 \cdot X^{k-1} + u_1 \cdot X^{k-2} + \cdots + u_{k-1}. \quad (1)$$

Let the parity check code be $(v_0, v_1, \ldots, v_{n-k-1})$, with $v(X) = X^{n-k}u(X) \mod g(X)$, then the code polynomial yields

$$c(X) = X^{n-k}u(X) + v_0 \cdot X^{n-k-1} + v_1 \cdot X^{n-k-2} + \cdots + v_{n-k-1}, \quad (2)$$

where $u(X) = \sum_{i=0}^{k-1} u_i \cdot X^{k-1-i}$, $n-k = 2t$, and t is the maximum number of error corrections. Then the encoded symbols will read $(v_{n-k-1}, \ldots, v_0, u_{k-1}, \ldots, u_0)$, with v_i and u_i being symbols in $GF(2^m)$.

Suppose the code generator is expressed as $g(X) = \prod_{i=1}^{2t}(X - \alpha^i) = g_0 + g_1 X + \cdots + g_{2t-1}X^{2t-1} + X^{2t}$. Then $g_0 = \prod_{i=1}^{2t} \alpha^i$, $g_1 = -\sum_{i=1}^{2t} \prod_{j \neq i}^{1 \leqslant j \leqslant 2t} \alpha^j$. For $2 \leqslant i \leqslant 2t-1$, the generator polynomial can be computed in Galois field recursively. In Matlab, the generator polynomial can be computed directly by function `rsgenpoly`.

Once the generator is calculated, the generator matrix can be obtained. Take $n = 255$, $k = 223$, $t = 16$, and $m = 8$, then the generating matrix can be computed in Matlab as follows:

```
n= 255;
k= 223;
t= 16;
m= 8;
tvc= zeros(1, 33);
tvc= [ 45  216  239  24  253  104  27 ...
       40  107  50  163  210  227  134 ...
       224  158  119  13  158  1  238 ...
       164  82  43  15  232  246  142 ...
       50  189  29  232  1];
gvc= gf(tvc, m);
dv= gf(zeros(k, n- k));
qv= gf(0, m);
gpc= fliplr(gvc);
 for i= 0: 1: k-1
     tmpv= gf([1, zeros(1, n- k+ i)], m);
     [Q, R]= deconv(tmpv, gpc);
     length1= length(R );
     indx= find(R~=gf(0, m), 1, 'first');
     dv(i+1, n- k- length1+ indx: n- k)
     = R(indx: length1);
 end
gtx= gf(zeros(k, n), m);
% gtx is the generating matrix
gtx(:, 1: n-k)= fliplr(dv);
gtx(:, n- k+ 1: n)= gf(eye(k, k), m);
```

The encoding of RS code can be computed as modular division in the polynomial field [14]. Let $b(X) = X^{n-k}u(X) \mod g(X)$, then $b(X) + X^{n-k}u(X)$ can be divided by $g(X)$ and become a RS code polynomial. The symbol u_0 with the lowest index is usually processed as the symbol with the highest weight for convenience.

The RS decoding is divided into four steps, i.e., syndrome computation, solution of error location polynomial, Chien search, and determination of the error values by Forney's algorithm. RS code is defined in the extension field $GF(2^{m \cdot s})$, and syndromes can be

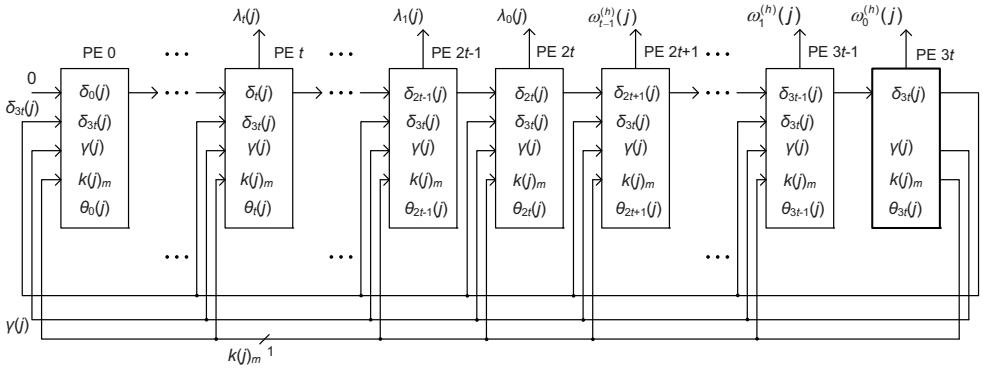


Figure 1: System architecture to solve error location polynomial [16, 1].

computed according to [15]:

$$S_i = r(\alpha^i) = \sum_{j=0}^{n-1} r_j \cdot \alpha^{i \cdot j}, \tag{3}$$

where $r_j \in GF(2^m)$, $1 \leqslant i \leqslant 2t$.

The error polynomial can be obtained by the improved inversion-less Berlekamp-Massey algorithm [16]. The resulting error polynomial reads

$$\sigma(X) = \sum_{i=0}^{e} \lambda_i X^i, \tag{4}$$

where $v \leqslant t$. λ_i is used to calculate the error locations. The error values are

$$\begin{aligned} Y_i &= \frac{z^{2t}\Omega^{(h)}(z)}{\sigma'(z)}\Big|_{z=X_i^{-1}}, \\ &= \frac{z^{2t} \cdot \sum_{i=0}^{e-1} \omega_i^{(h)} \cdot z^i}{\sum_{i=2k-1}^{k=1,\ldots,\lceil \frac{e}{2} \rceil} \lambda_i \cdot z^{i-1}}\Big|_{z=X_i^{-1}}, \end{aligned} \tag{5}$$

where $\Omega^{(h)}(z) = \sum_{i=0}^{e-1} \omega_i^{(h)}(z) \cdot z^i$, the symbol h denotes the higher partial products from z^{2t} to z^{2t+e-1}, and e is the number of error locations.

The system architecture of RS decoding [16, 1] is shown in Fig. 2. First, the original passages through processing elements are replaced by direct connections of signals. Second, the control logics are totaly integrated into the last processing element for simplicity. The signal $k(j)_m$ is the sign bit of the integer $k(j)$, where m denotes the m-th bit. The coefficients from λ_t to λ_0 come out of from PE t to PE $2t$, while the coefficients from $\omega_{t-1}^{(h)}$ to $\omega_0^{(h)}$ are produced from PE $2t+1$ to PE $3t$.

Among the processing elements, there exists matrix products. Typically, the matrix multiplication $[\alpha^m, \alpha^{m+1}, \ldots, \alpha^{2m-2}]^T = Q \cdot [1, \alpha, \ldots, \alpha^{m-1}]^T$, the $GF(2^m)$ can be computed according to $C = A \cdot B = (L + Q^T \cdot U) \cdot B$ proposed in [17]. In Matlab, the matrix Q in $GF(2^8)$ can be obtained in the following:

```
Tx= zeros(m-1, m);
Ty1= zeros(m, m);
vf0= [1, 0, 1, 1, 1, 0, 0, 0];
Ty(:, 1: m-1)= [zeros(1, m-1); eye(m-1)];
Ty(:, m)= vf0.';
Ty1= Ty^m;
Ty1= mod(Ty1, 2);
Tx= Ty1(:, 1: m-1).';
Q= Tx;
```

The Chien search circuit is composed of t multipliers in $GF(2^m)$ and a few adders. It is used to find out the error locations by a full search if $\lambda_i \neq 0$.

3 Related Work

The popular stream cipher consists of two key stream generators, as is shown in Fig. 2 [18, 19]. The plaintexts are masked by the keys on the left, and then decrypted by the same stream on the right. The symbol 'IV' denotes initialization vector.

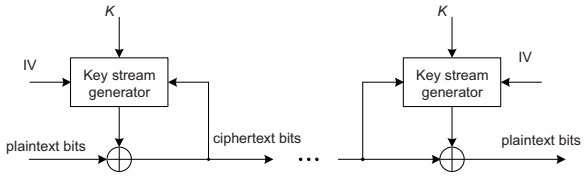


Figure 2: Traditional stream cipher by two key generators.

A random number generator (RNG) can be implemented by the linear feedback shift register (LFSR)[20] that generates peudo-random numbers for stream cipher, as is shown in Fig. 3. In practise the RNG can also be constructed by other digital or analog circuits.

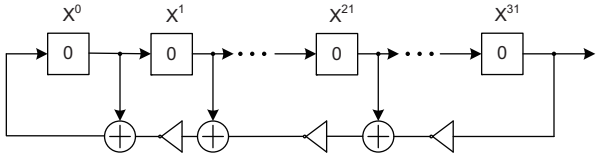


Figure 3: Linear feedback shift registers by XNOR gates[21].

A light-weight and energy-friendly stream cipher is proposed for wireless sensor network (WSN) in [18]. The paper suggests use the linear combination of pervious packets to generate the peudo-random key bits in a stream cipher, where a mixed protocol is advanced and public key is used to achieve data privacy [18].

Joint encoding and encryption by LDPC codes and RS codes for public key cryptography is a tradeoff between security and reliable communication [5, 22, 6]. New scrambling and permutating matrices for McEliece system with RS codes are proposed in [6] to enhance security. In [2] the wire-tap encoding and the error correction encoding are concatenated to enhance security. And in [3], a generator matrix, an invertible matrix, and gamma generator are used to produce the ciphertexts.

A provable secure stream cipher based syndrome decoding problem is presented in [7], which uses regular words to speed the system up and quasi-cyclic codes to reduce memory requirements. The results seem to be based on quasi-cyclic BCH codes.

4 Proposed Approach

Now suppose the message piece is r and the random numbers s are combined or XORed, then one can encode $(r \oplus \alpha \cdot s)$, where α is the multiple of s or a $u \times v$ matrix with $u = s/w$, $v = r/w$ and w is the width of a symbol. Next, one piece of the symbols s is inserted by addition modulo 2^m, and the new message reads $r' = M + s$. At the other terminal, by RS decoding one can recover s as well as $(r + \alpha s)$. Finally, the $r = (r + \alpha s) - \alpha s$ can be restored. The whole process is shown in Fig. 4. Especially, the insertion of errors avoids zero bytes by changing them as nonzero bytes within random numbers.

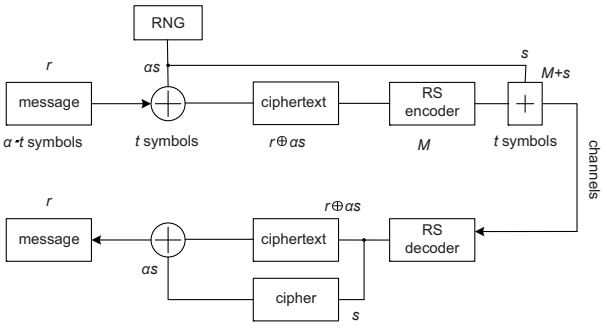


Figure 4: Stream cipher by RS code.

First, the security of a code-based system can be seemed as a public-key security problem, while the usual stream cipher is a symmetric cryptosystem. The proposed mechanism embeds the stream cipher along Reed-Solomon codes looking like a public-key cryptography: considering the key streams as private keys for encoding, and looking at the RS code parameters as the public key to decrypt the ciphertexts.

The security of the proposed scheme may also be improved by the application of shortened codes. In this way, the n-symbol codewords are replaced by a $(n - \delta)$-symbol codewords. Take the wireless communication for example, the sender and receiver share the same hardware units, i.e., stream engines, encoders and decoders. Then, the stream cipher with the sender can be seen as private keys to encrypt the messages. The choices of (1) whether a subfield subcode is chosen (2) whether a shortened code is chosen (3) whether the same decoder exists can be used as the public keys. If the decoder exists and the other two parameters are rightly chosen, then the stream cipher can be separated from the message bits by RS decoding.

In detail, one can use subcode or shortened code to decrease the error rate or improve the security of the system, as is shown in Fig. 5. The subfield subcode over $GF(p^s)$ have the same length n but is smaller than its parent code [23] with the word width, which is compressed from $\lceil sm \log_2 p \rceil$ to $\lceil s \log_2 p \rceil$. In fact, the length n of a subfield subcode may be larger than the size of a subfield, but is constrained its parent field size. Notice that the parity check symbols keep about double word size of other code symbols in Fig. 5.

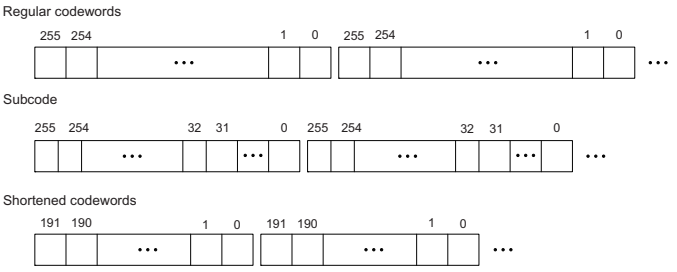


Figure 5: Stream cipher by subcode or shortened RS codes.

Secondly, the proposed scheme reduce the hardware cost and complexity by reusing the encoder for encryption and recovering the key streams by the decoder. Therefore, the task of aligning the stream engines to time sequence becomes unnecessary. As the number of stream cipher engines decreases from 2 to 1, it reduces the continuous power consumption much. On the one hand, it reduces the hardware units by a half; on the other hand, the renewal frequency of the stream cipher engine can be lowered down. Also, it is not easy to align the messages to real time, but it is feasible to align the messages at code sequence. The oscillator frequency may vary from time to time, while the stream cipher engines requires accurate alignments. The existence of time delay in mobile phones proves the validity of code alignment.

Furthermore, if the function of RS decoding and encoding is occasionally used, then the encryption can be used as an auxiliary function. Otherwise, if the error correction through the transmission path is requirable, then the first few code symbols can be designated as masks of key streams while the next words can be corrected by the remained error correction capability. Since error correction logic and stream cipher are basic units in wireless communications, this scheme

secret data as stream cipher

```
00000000h: 31 30 38 3D 33 5E 33 2A 32 5E 32 0D 0A 0D 0A  ; 108=3^3*2^2...
```

message

```
00000000h: 54 48 65 20 57 61 74 65 72 20 4D 61 72 67 69 6E ; THe Water Margin
00000010h: 20 74 65 6C 6C 73 20 74 68 65 20 73 74 6F 72 79 ;  tells the story
00000020h: 20 6F 66 20 72 65 62 65 6C 6C 69 6F 6E 20 61 6E ;  of rebellion an
00000030h: 64 20 73 74 72 75 67 67 6C 65 20 6F 66 20 73 6B ; d struggle of sk
00000040h: 69 6C 6C 65 64 2C 20 62 72 75 74 65 2C 20 6F 72 ; illed, brute, or
00000050h: 20 62 72 61 76 65 20 6D 65 6E 20 69 6E 20 61 63 ;  brave men in ac
00000060h: 69 65 6E 74 20 43 68 69 6E 61 2E 20 54 68 65 79 ; ient China. They
00000070h: 20 61 72 65 20 6E 61 6D 65 64 20 4A 69 61 6E 67 ;  are named Jiang
00000080h: 20 53 6F 6E 67 2C 20 4A 75 6E 79 69 20 4C 75 2C ;  Song, Junyi Lu,
00000090h: 20 53 68 65 6E 67 20 47 6F 6E 67 73 75 6E 2E 2E ;  Sheng Gongsun..
000000a0h: 2E 0D 0A 0D 0A 0D 0A                           ; .......
```

message encrypted by stream cipher

```
00000000h: 65 78 5d 1d 64 3f 47 4f 40 7e 7f 6c 78 6a 63 6e ; ex]d?GO@~lxjcn
00000010h: 11 44 5d 51 5f 2d 13 5e 5a 3b 12 7e 7e 62 78 79 ; □D]Q_-□^Z;□~~bxy
… …
```

Figure 6: Hiding information as stream cipher in RS codes.

Table 1: Advantages of Proposed Stream Cipher Compared to Other Schemes

Reference	Function	Implement	Topology	ENC/DEC Symmetry	Reuse for ECC	No. of RNG	Randomness	Code
This work	Stream data	Hardware	1-to-N	Not full	Yes	1	Not strictly	RS
[10]	Stream data	Hardware	1-to-1	Symmetric	No	2	Required	No
[7]	Stream data	Software	1-to-1	Symmetric	No	2	Required	Quasi-cyclic
[8]	Exchange data	Software	N-to-N	Not	No	0	-	GRS

combines their functions.

Thirdly, it may be applicable to transmit extra data through the key streams and therefore increase the throughput of inserted codewords. For example, the $(255, 223, 16)$ RS code has 223 information symbols. If the encoder are used for stream cipher, then 16 symbols of information can be added to the codeword secretly. The information rate is increased from $223/255 = 0.87$ to $239/255 = 0.94$. As is shown in Fig. 6, it is able to mix information with secret data in the proposed codes.

5 Comparison

The comparisons of this work compared with peer works are shown in Tab. 1. It should be noticed that the proposed scheme is not full symmetric, for which the cipher is decoded through the public-key. The architecture is implemented in a topology of 1-to-N, with N being any natural numbers including 1. If $N = 1$, then there are two parties to communicate with each other; or else it can be used in a network. By contrast, the stream cipher is usually supposed to follow 1-to-1 topology. While the encoding and decoding architectures in our work are non-symmetric, the whole stream cipher engine can be made nearly symmetric by integrating an encoder and decoder together. Moreover, this work needs only 1 random number generator while other stream ciphers requires 2 to N random number generators, since the key stream is embedded in the data stream in this work.

The weakly secure data exchange problem with generalized Reed-Solomon code is discussed in [8], in which data exchange between multiple parts rather than two parts are considered.

Also, in the proposal the random number generator is not that strictly required and pseudo random numbers can be used, while other works pay much attention to the random number generator itself, owing to different mechanism for stream ciphers. While traditional stream cipher uses continuous random bits, this work applies data structures of codewords to separate data. The references [8, 7] also uses GRS codes and quasi-cyclic codes to encode data and bring out randomness.

The architecture and hardware implementation results with [10] is demonstrated in Tab. 1 and Tab. 2. It uses a hardware common key cryptography named RAC for stream cipher, where the random numbers are used as addresses to relocate the data bits. The random numbers are generated by the recipient as addresses to resort the received ciphertext. Assuming the stream cipher engine in [10] works in a pipeline, then it is able to encrypt and decrypt 1 byte every clock cycle with throughput up to 3.2 Gps. Nevertheless, considering that the data from the sender or to the recipient should be stored in a RAM for reorganization, so the real decryption throughput may be only a half, let alone the time for fetching instructions.

Finally, the power consciousness of this design is

demonstrated by the application of RS code, whereas McEliece systems usually use nonsystematic encoding with an increase of computational efforts by $n/(n-k)$ times. In addition, if there are demands of error corrections for channels, the hardware units may be reconfigured and reused at different times.

Table 2: Hardware Implementation of Stream Cipher Engine by Reed-Solomon Code.

Design	Platform	LUTs	f_{max} (MHz)	Throughput (Gps)
Enc	XC5VLX85-2	318	480	3.79
Dec	XC5VLX85-2	5253	270	0.289
Enc/Dec	XC5VLX85-2	5905	270	1.82/0.289
[10]	0.18 μm CMOS	1.38 mm^2	400	1.6∼ 3.2

The stream cipher with encoding and decoding by RS codes is described by Matlab for test at first, then it is described by Verilog HDL for hardware implementation. As is shown in Tab. 2, the stream cipher engine is simulated by Modelsim 6.2, synthesized by Synopsys Synplify Pro 2014, and placed and rooted in Xilinx ISE 14.7. The code parameters are chosen as usual in [24], i.e.,
$m = 8$: number of bits per symbol,
$n = 255$: number of symbols per codeword,
$k = 223$: number of message symbols in a codeword,
$t = 16$: number of corrected errors by symbols.

According to the theory and results in [7], the security level of the cipher in the above table is about *80* bits of keys with symmetric cryptography.

In Tab. 2, the RS code supports correcting 16 symbols of errors along with 223 symbols of messages. It takes about 3.78 μs to correct 2 errors of symbols, and 7.06 μs to correct 16 errors of symbols. Its information throughput without parity check bits are 1.59/0.253 Gbit/s respectively for encoding and decoding.

6 Conclusion

Stream cipher is often used for mobile communications between two parties. In general, this paper proposes an easy scheme to implement stream cipher along with the popular RS code, which relies on RS encoding and decoding to cover and uncover the key streams. It is beneficial to easily evaluate the complexity and enhance the security of communication systems. The proposed cipher is based on the NP-complete problem of decoding a random linear code, which simply conforms to many solved code-based security issues. The cipher also decreases one key generator and reduces unnecessary power consumption for the synchronization between parties.

Acknowledgment

The author would like to appreciate the editor and reviewers for their kind help. This work is supported by Shenzhen postdoctoral base for innovation and practice and partly by Shenzhen basic research program (No. JCYJ20170307141601162).

References

[1] T. Wu and R. Wang, "Stream cipher by Reed-Solomon code," in *International Conference on Information and Communication Technology Convergence (ICTC)*, 2017, pp. 422–427.

[2] F. E. Oggier and M. J. Mihaljevic, "An information-theoretic analysis of the security of communication systems employing the encoding-encryption paradigm," arXiv, 2018. [Online]. Available: http://arxiv.org/abs/1008.0968

[3] A. N. Alekseychuk and S. V. Gryshakov, "Secure and practical randomized stream ciphers based on ReedSolomon codes," *Cybernetics and Systems Analysis*, vol. 53, no. 2, pp. 262–268, 2017.

[4] S. Lin and J. Daniel J. Costello, *Error Control Coding*. Upper Saddle River, New Jersey: Pearson Education, Inc., 2004.

[5] M. Baldi, *QC-LDPC Code-Based Cryptography*, ser. Briefs in Electrical and Computer Engineering. Springer, 2014.

[6] M. Baldi, F. Chiaraluce, M. Bianchi, and J. R. D. Schipani, "Enhanced public key security for the McEliece cryptosystem," *Journal of Cryptology*, vol. 29, no. 1, p. 127, January 2016.

[7] P. Gaborit, C. Lauradoux, and N. Sendrier, "SYND: a fast code-based stream cipher with a security reduction," in *IEEE International Symposium on Information Theory*, 2007, pp. 186–190.

[8] M. Yan, A. Sprintson, and I. Zelenkoy, "Weakly secure data exchange with generalized reed solomon codes," in *IEEE International Symposium on Information Theory*, 2014, pp. 1366–1370.

[9] M. Orumiehchiha, J. Pieprzyk, E. Shakour, and R. Steinfeld, "Security evaluation of Rakaposhi stream cipher," in *International Conference on Information Security Practice and Experience (ISPEC 2013)*, ser. Lecture Notes in Computer Science, R. Deng and T. Feng, Eds., vol. 7863. Berlin, Heidelberg: Springer, 2013, pp. 361–371.

[10] M. Fukase, H. Takeda, R. Tenma, K. Noda, Y. Sato, R. Sato, and T. Satot, "Development of a multimedia stream cipher engine," in *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS2006)*, Tottori, Japan, 2006, pp. 562–565.

[11] H. C. van Tilborg, *Fundamentals of cryptology: A Professional Reference and Interactive Tutorial*. Boston/Dordrecht/London: Kluwer Academic Publishers, 1999.

[12] T. Eisenbarth, T. Guneysu, S. Heyse, and C. Paar, "Microeliece: Mceliece for embedded devices," in *CHES*, ser. LNCS, vol. 5747, 2009, p. 4964.

[13] S. Heyse, I. von Maurich, and T. Güneysu, "Smaller keys for code-based cryptography: QC-MDPC McEliece implementations on embedded devices," in *CHES*, ser. LNCS, vol. 8086, 2013, pp. 273–292.

[14] S. Lin and J. Daniel J. Costello, *Error Control Coding- Fundamentals and Applications*. New Jersey: Prentice Hall, 1983.

[15] B. Sklar, "Reed-solomon codes," 2001. [Online]. Available: http://ptgmedia.pearsoncmg.com/images/art_sklar7_reed-solomon/elementLinks/art_sklar7_reed-solomon.pdf

[16] D. V. Sarwate and N. R. Shanbhag, "High-speed architectures for reedsolomon decoders," *IEEE Transactions on VLSI Systems*, vol. 9, no. 5, pp. 641–655, October 2001.

[17] A. Reyhani-Masoleh and M. A. Hasan, "Low complexity bit parallel architectures for polynomial basis multiplication over $GF(2^m)$," *IEEE Transactions on Computers*, pp. 945–959, 2004.

[18] H. Lin, A. J. A., D. Bai, and Y. Liu, "A light-weight stream cipher for wireless sensor networks," in *Future Wireless Networks and Information Systems*, ser. Lecture Notes in Electrical Engineering, Y. Zhang, Ed., vol. 143. Berlin, Heidelberg: Springer, 2012, pp. 627–638.

[19] B. Sklar, *Digital Communications Fundamentals and Applications*. Beijing: Publishing House of Electronics Industry, 2015, in Chinese.

[20] A. Canteaut, "Linear feedback shift register," pp. 726–729, 2011, reference Work Entry.

[21] "Linear feedback shift registers in virtex devices," 1996. [Online]. Available: http://www.xilinx.com/support/documentation/application_notes/xapp052.pdf

[22] M. Esmaeili and T. A. Gulliver, "Joint channel coding-cryptography based on random insertions and deletions in quasi-cyclic-low-density parity check codes," *IET Communications*, vol. 9, no. 12, pp. 1555–1560, 2015.

[23] J. I. Hall, http://users.math.msu.edu/users/jhall/classes/codenotes/Subfields.pdf, Michigan State University, pp. 89–99, 2012, codes over Subfields.

[24] Z. Liang and W. Zhang, "Efficient Berlekamp-Massey algorithm and architecture for Reed-Solomon decoder," *Journal of Signal Processing Systems*, vol. 86, p. 5165, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 30-37 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Performance improvement of a wind energy system using fuzzy logic based pitch angle control

Kanasottu Anil Naik*, Chandra Prakash Gupta, Eugene Fernandez

Department of Electrical Engineering, IIT Roorkee, 24767, India

ARTICLE INFO

Article history:
Received: 21 March, 2018
Accepted: 20 April, 2018
Online: 07 May, 2018

Keywords :
Fuzzy logic controller
Pitch Angle controller
Transient Stability
Wind Energy System

ABSTRACT

The pitch angle controller maintains the aerodynamic captured power at rated level when the wind speed is above the rated speed. Besides, it can also improve the transient stability occurring in the wind energy system (WES). This paper, therefore, proposes an effective pitch angle control strategy that can deliver the conditioned output power in windy condition and increase the transient stability capability in grid faults. A fuzzy logic method has employed to design the proposed control strategy, Moreover, in this paper some major factors that affect the transient stability have been investigated by deriving steady-state equivalent model of the wind energy system. The simulated results show that the proposed fuzzy logic based pitch angle controller is effective at conditioning the output power and complying with fault ride through requirements for WES in the power system.

1. Introduction

Transient stability enhancement of wind energy system using fuzzy logic based pitch angle controller is a topic of much interest in wind energy studies [1]. Wind energy is perhaps more economical and viable among the other renewable energy resources available for electric power generation. According to predictions it is estimated that, by 2020 wind energy will contributes to about 10 percentage of the world's electricity [2]. While wind energy is free, its generation involves some operational issues. The main issue that arises is that wind speed is not constant, resulting in large fluctuations in the real power since the output power is proportional to the cube of the wind speed. Another issue of concern is that if a fault occurs in the power network then the induction generator that is generally used in a wind energy system (WES) will not be supplied with the necessary reactive power and this will result in rotor instability [3].

Pitch angle controller limit is one of the ways to regulate the speed of the wind turbine under varying wind speeds. This controller helps to maintain the aerodynamic power at rated value when the wind speed exceeds the rated wind speed. In some reported works [4-9], different methods have been proposed to design the pitch angle controller for regulating the active power at rated value and improve the power quality with frequent wind speed changes. Beside this, pitch angle controller is also employed to stabilize the WES under transient disturbances or faults. The transient stability analysis of a WES is important because of its impact on the utility grid. In the literature some researchers have designed rotor speed control based pitch angle controller to enhance the transient stability [10-13].

Thus, the pitch angle control serves two purposes: In [4-9], they are used to regulate the generator output power at rated value for varying wind speeds while in [10-13], they are used for stabilizing the WES under transient disturbances or post fault operation. A study of the performance of the pitch controller in both these conditions is the subject of this paper. A pitch angle control strategy is proposed which is designed for improving the smoothing of power and transient stability behavior of the WES under disturbances or wind speed variations. Fuzzy logic technique has been employed for the pitch-angle controller design, as it reported best suited in comparison to other types of controllers.

The major contribution of this paper as follows:

- Derived a steady-state equivalent model of the induction generator using analytical approach. The mechanical torque versus rotor speed results were obtained under different pitch angle conditions.
- A fuzzy logic controller (FLC) based pitch angle control system (in power and speed control modes) has proposed, which can enhance the transient stability performance of the WES subjected to severe network disturbances (in speed control mode) and maintain the output power when the wind speed is higher than rated speed (in power control mode).

This paper is structured as follows: Section 2 discusses the system configuration, modeling of wind turbine and induction generator. In section 3, Pitch angle controller basic concept, proposed control strategy, and controller design (PI and fuzzy

*Kanasottu Anil Naik, EED, IIT Roorkee, India, +919690314482,
Email: anilnaik205@gmail.com

logic) are discussed in detailed. The simulation results are discussed in section 4. Finally, important conclusions are drawn in section 5.

2. Modelling of Wind Energy System

2.1. Configuration of the studied system

A typical WES that is used for the present simulations consists of a 1.5 MW wind turbine driven induction generator (IG) as shown in Fig. 1. The stator winding of the IG is connected to the point of common coupling (PCC) through a step-up transformer (0.69/25kV) which exports power to the 120kV grid through step-up transformer (25/120kV) and a transmission line ($R_{TL} + jX_{TL}$) operating at 120kV. The rotor of the IG is driven by a variable pitch wind turbine. A power factor correction capacitor (C) is connected to the low voltage terminal of the wind turbine generator. It provides required amount of reactive power to the IG for maintaining the terminal voltage. The design based parameters of the generator and wind turbine are given in Appendix-I.

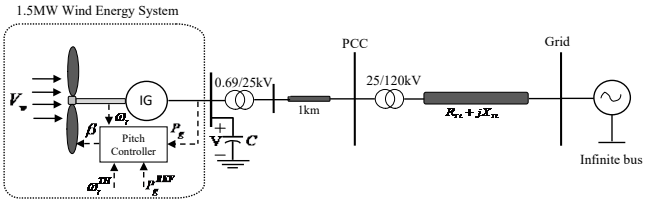


Figure. 1. The studied grid-connected WES

2.2. Wind Turbine Modeling

The developed mechanical power of a typical wind turbine is directly proportional to the cube of wind speed as given:

$$P_m = \frac{1}{2}\rho A_r C_p(\lambda,\beta)V_w^3 = C_p(\lambda,\beta)P_w \quad (1)$$

where, P_w is available wind power, ρ is the air-density (kg/m^3), A_r is the turbine swept area (m^2) and can be written as $A_r = \pi R^2$, V_w is the wind speed (m/s) and C_p is the power coefficient.

The typical wind turbine is characterized by the power coefficient (C_p) which depends upon the ratio of rotor-tip speed (λ) and blade pitch-angle (β). The power coefficient (C_p) employed in this study is as follows [14]:

$$C_p(\lambda,\beta) = c_6\lambda + e^{-c_5/\lambda_i}\left(-c_4 - c_3\beta + c_2/\lambda_i\right)c_1 \quad (2)$$

where

$$\frac{1}{\lambda_i} = \frac{1}{0.008\beta + \lambda} - \frac{0.035}{1+\beta^3} \quad (3)$$

The coefficients $c_6 - c_1$ are: $c_6 = 0.0068$, $c_5 = 21$, $c_4 = 5$, $c_3 = 0.4$, $c_2 = 116$ and $c_1 = 0.5176$ and the tip-speed ratio (λ) can be defined as:

$$\lambda = \frac{\text{tip speed of the blade}}{\text{wind speed}} = \frac{\omega_r R}{V_w} \quad (4)$$

where, R is the radius of turbine rotor [m] and ω_r is rotor speed [rad/s]

The mechanical power (P_m) shown in equation (1) can be expressed as:

$$P_m = \frac{1}{2}\rho\pi R^2 C_p(\lambda,\beta)V_w^3 = \frac{1}{2}\rho\pi R^2 C_p(\lambda,\beta)\omega_r^3\left(\frac{R}{\lambda}\right)^3 = C_p(\lambda,\beta)K\omega_r^3 \quad (5)$$

where, $K = \frac{1}{2}\rho\pi R^2\left(\frac{R}{\lambda}\right)^3$

Hence, the mechanical torque output of induction generator (IG) can be derived as [15]:

$$T_m = \frac{P_m}{\omega_r} = \frac{C_p(\lambda,\beta)P_w}{\omega_r} \quad (6)$$

According to equations (1) – (4), for different values of pitch-angle (β), the power coefficient versus tip-speed ratio ($C_p - \lambda$) curve for the studied system has been obtained as shown in Fig. 2. For different values of wind speed (V_w), the turbine output power versus turbine speed characteristics is as shown in Fig. 3.

The relation between mechanical torque (T_m) and pitch angle (β) with respect to rotor speed (ω_r) as shown in Fig. 4 is obtained using information of Fig. 2 with Eqn. (6). As observed from equation (6), the mechanical power output (P_m) of wind turbine depends upon power coefficient (C_p) and the rotor speed (ω_r) of wind turbine. However, the turbine speed varies very little as the fixed speed cage generators have a speed variation range less than 1% [16]. Thus, variation of mechanical power depends mainly on the power coefficient $C_p(\lambda,\beta)$ which varies with the tip speed ratio (λ) and pitch angle β . However, it may be observed that in transient disturbances, tip speed ratio variation is very small [17]. As a result, the mechanical torque mostly depends on the pitch angle (β). So, it can be noticed from Fig. 4, that the mechanical torque decreases, at the particular point onwards of rotor speed, as the pitch angle increases.

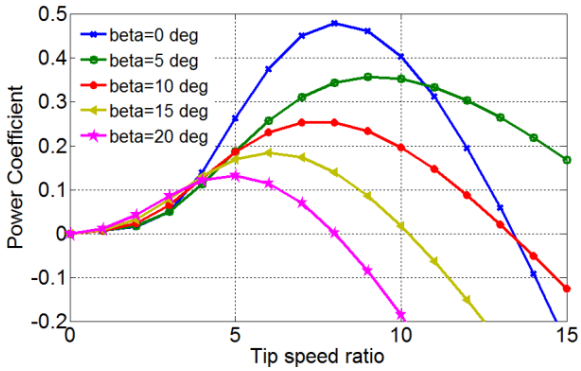


Figure. 2. Wind turbine $C_p - \lambda$ curve

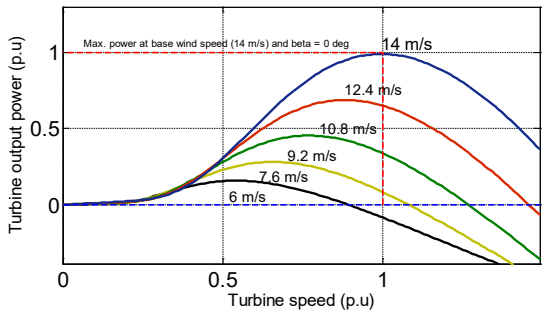


Figure. 3. Turbine Power characteristics

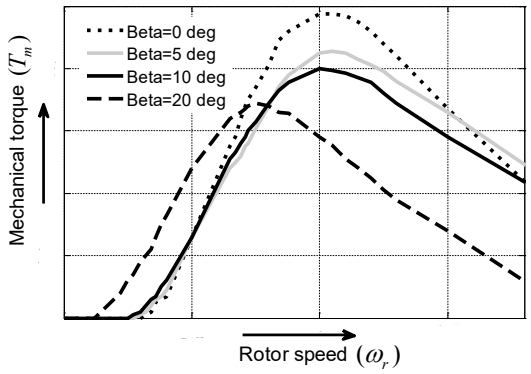


Figure. 4. Mechanical torque versus rotor speed

2.3. Induction generator modeling

The employed test system shown in Fig. 1 can be modeled by the equivalent electrical circuit from induction generator (IG) to PCC is shown in Fig. 5. The steady state equivalent circuit of the IG is as shown Fig. 5 within the box enclosed by dotted lines.

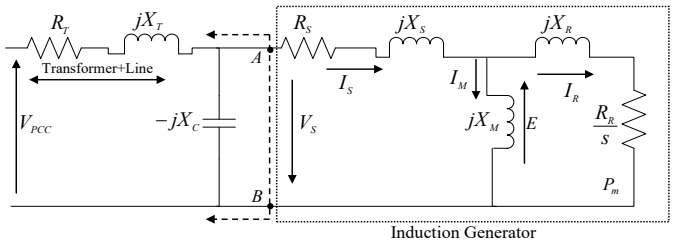


Figure. 5. Electrical equivalent of the test system

According to the IG equivalent circuit, the mathematical expression for stator terminal voltage (V_S) of the IG is:

$$V_S = I_S(R_S + jX_S) + E \tag{7}$$

where, E is air gap magnetic field induction electro-magnetic-field (EMF) given by

$$E = jI_M X_M \tag{8}$$

Where X_M magnetizing reactance and I_M is the exciting current of the IG. Now the rotor current I_R can be determined from stator current I_S after deducting I_M as follows:

$$I_R = I_S - I_M \tag{9}$$

and then the corresponding mechanical power input to the IG will be given by

$$P_m = 3I_R^2\left(\frac{R_R}{s}\right) \tag{10}$$

where, R_R and R_S are the rotor and stator resistance, respectively and s is the slip which is less than zero for the IG. To understand the transient stability concept of the studied wind power system it is necessary to calculate the electrical torque of the generator using the electrical equivalent circuit of the system. The Thevenin equivalent circuit obtained from Fig. 5 from the point A and B is as shown in Fig. 6.

Where, Thevenin equivalent voltage is:

$$V_{1TH} = \frac{V_{PCC}(-jX_C)}{R_T + jX_T - jX_C} \tag{11}$$

and the series parameters are:

$$R_{1TH} = R_T \text{ and } X_{1TH} = X_T - X_C$$

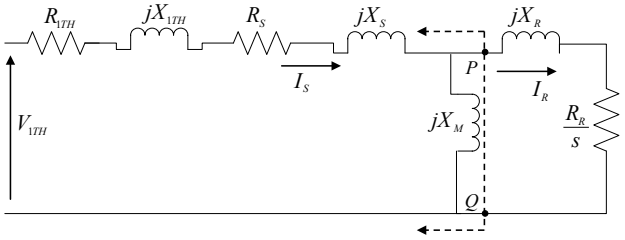


Figure. 6. Reduced Equivalent circuit of the system

Again, applying the Thevenin concept to Fig. 6 at the point P and Q the further reduced equivalent circuit has been obtained as shown in Fig.8. The Thevenin equivalent voltage obtained as:

$$V_{TH} = \frac{V_{1TH}(jX_M)}{R_{1TH} + R_S + jX_{1TH} + jX_S + jX_M} \tag{12}$$

and the equivalent Thevenin resistance and reactance can be obtained as: $R_{TH} = R_{1TH} + R_S$ and $X_{TH} = X_{1TH} + X_S + X_M$.

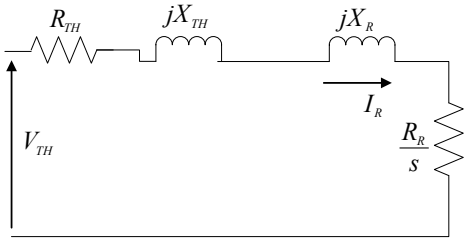


Figure. 7. Complete Reduced Equivalent circuit of the system

Finally, from the Fig.7, the rotor current can be computed as [18]:

$$I_R = \frac{V_{TH}}{\sqrt{\left(R_{TH} + R_R / s\right)^2 + \left(X_{TH} + X_R\right)^2}} \tag{13}$$

From which, the generator electrical toque can be determined as:

$$T_e = \frac{R_R}{s} I_R^2 = \frac{R_R}{s} \frac{V_{TH}^2}{\left(R_{TH} + R_R / s\right)^2 + \left(X_{TH} + X_R\right)^2} \tag{14}$$

Generally, the slip (s) is defined as:

$$s = \frac{\omega_s - \omega_r}{\omega_s} \tag{15}$$

When the induction machine is operated as generator, the sign convention for the electrical (T_e) and mechanical (T_m) torques are reversed. Therefore, the mechanical-electrical equilibrium of the IG can be expressed as

$$\frac{d\omega_r}{dt} = \frac{1}{2H}\left(T_e - T_m\right) \tag{16}$$

Eqn. (16) gives the information that during the network faults the electrical torque (T_e) become zero since the voltage at the PCC is zero which causes the generator rotor speed to increase drastically leads to instability of the system. However, by increasing the pitch angle of wind turbine the mechanical toque (T_m) can be decreased (see Fig. 4) and thus, generator rotor speed (ω_r) can be limited/decreased to its normal value as given by Eqn. (16). Thus, the pitch angle controller plays an important role in the operation of a WES to improve the transient stability.

3. Pitch angle controller

The block diagram of a typical pitch-angle control system is shown in Fig. 8. When the wind speed crosses the rated wind speed (V_{wR}), the pitch angle controller modifies the aerodynamic toque to maintain the generator output power at its rated value. The difference between the measured power (P_g) and reference power (P_g^{REF}) generates a control signal to the controller $C(s)$, which regulates the output power accordance with error (ε), by modifying the pitch angle.

The pitch-angle control strategy mathematically can be expressed as:

$$\beta_c = \frac{\Delta\beta}{\Delta P}\left(P_g - P_g^{REF}\right) + \beta_i \tag{17}$$

where, ΔP and $\Delta\beta$ are small-signal state variables of mechanical power and pitch angle, respectively, and β_i is the initial pitch-angle magnitude.

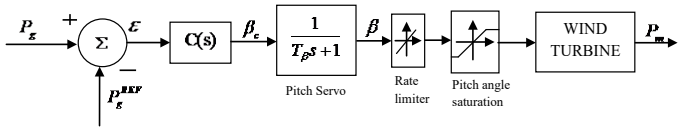


Figure. 8. The typical pitch-angle control system

The optimum pitch angle (β_c) from the controller $C(s)$ is used as reference to the pitch actuator/servo system. The purpose of the pitch servo is for the correct positioning of the blades. The first order transfer function of the electric or hydraulic pitch servo is as follows:

$$\beta = \frac{1}{T_\beta s+1}\beta_c \tag{18}$$

The pitch angle (β) follows the reference pitch or optimum pitch (β_c) by a first order lag with time constant T_β, which is dependent on the pitch actuator. In order to get the realistic response from the pitch control system, the pitch rate and the regulations range of pitch angle are set to $\pm 2°/\text{sec}$ and $0° - 45°$, respectively.

As discussed earlier, the pitch angle controller is generally employed for control above rated wind speed where it regulates the generator output power. Besides, it can also be used to improve the transient stability of the WES. The next subsection discusses how the pitch angle controller can serve these tasks well.

3.1. Control Strategy implemented

The controller shown in Fig. 9 can be operated in two control modes (power and speed) in order to achieve above tasks effectively. The operating mode of the controller can be determined using selection switch. The controller works in power control mode with switch set to input 1 when the wind speed exceeds its rated speed to maintain the output power of the generator at rated value. However, for below rated wind speed controller not in operation where $\beta = 0°$ to extract the possible maximum wind power as governed by the Fig. 2. On the other hand, if the rotor speeds of induction generator increase to a threshold rotor speed (ω_r^{TH}) due to network disturbances, the controller input is set to input 2, as a result, it works in speed control mode where the transient stability enhancement is achieved with suitable pitch angle generation.

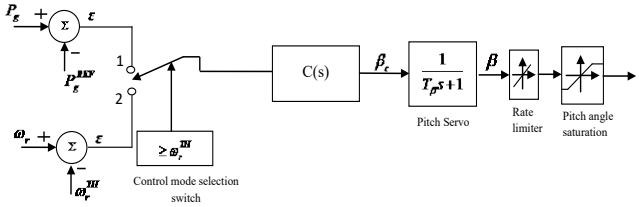


Figure.9. A typical pitch angle controller in power and speed control modes

In order to achieve the effective performance of the wind energy system, the controller C(s) of pitch control system needs to be designed for effective control.

3.2. Proportional Integral (PI) controller

A PI controller was designed using the linearized model of the WES. The Ziegler-Nichols method is employed to determine the appropriate controller gain. When the wind speed is above the rated wind speed for the WES, the PI controller is used to regulate the output power at its rated value by setting a suitable valued pitch angle. Although, this method has a simple structure, good steady-state response and is easy to implement, it exhibits a poor transient response due to use of constant controller gains. Therefore, an alternative method for controller design is proposed using fuzzy logic. It has numerous advantages as fuzzy logic does not require a mathematical modeling of the system and is capable of handling irregularity /nonlinearity. In addition, they provide excellent dynamic responses.

3.3. Design of Fuzzy logic controller

In 1965, Professor Lofti Zadeh proposed the concept of fuzzy logic (FL) [19]. The aim of FL was to develop an output by allowing a set of membership functions(MF) to decide the value of the inputs rather than representation using crisp deterministic values. Fuzzy based controllers have been found to be excellent choices for many control applications, as they imitate human control logic closely. The fuzzy rules are framed on the basis of the expert knowledge gained by the study of the performance of the system over a period of time. The expert knowledge so gained about system output behavior is expressed in terms of membership functions of various control parameters. The block diagram of a conventional FLC (named as Type-1 FLC) is as shown in Fig. 10. It consists of fuzzifier, inference engine and defuzzifier. Here, the fuzzifier converts the crisp value of the input parameter into fuzzy set and depending upon the fuzzy rules framed based on the membership functions for a given parameter of the system and then, fuzzy outputs are obtained with the help of the inference engine. Finally, a defuzzifier converts these fuzzy inputs to a crisp output of the FLC to be used for the control purpose. The employed fuzzy logic based pitch angle controller is shown in Fig. 11.

A. Fuzzification

The input and output variables used for the controller design are expressed with the help of fuzzy sets using linguistic variables. The designed MFs for FLC are shown in Fig. 12. The fuzzy sets for power error ($\mathcal{E}$) have been defined as: N(Negative), ZE (Zero), XSP(Extra Small Positive), SP(Small Positive), MP(Medium Positive), LP(Large Positive), and XLP(Extra Large Positive). For the change-in-power error ($\mathcal{CE}$) the fuzzy sets are chosen as NB (Negative Big), NS (Negative Small), ZE (Zero), PS (Positive Small), and PB (Positive Big). While for the output pitch-angle control, six membership functions have been defined as ZE (Zero), XS (Extra Small), S (Small), M (Medium), L (Large), and XL (Extra Large) [20,21].

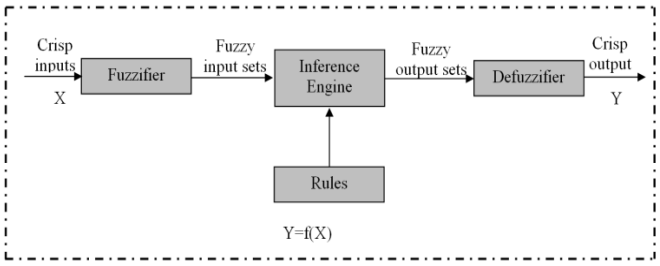


Figure. 10. Schematic diagram of classical (Type-1) Fuzzy Logic Controller

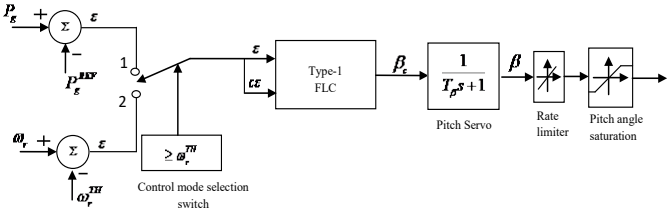


Figure. 11. Fuzzy logic (Type-1) based pitch angle controller

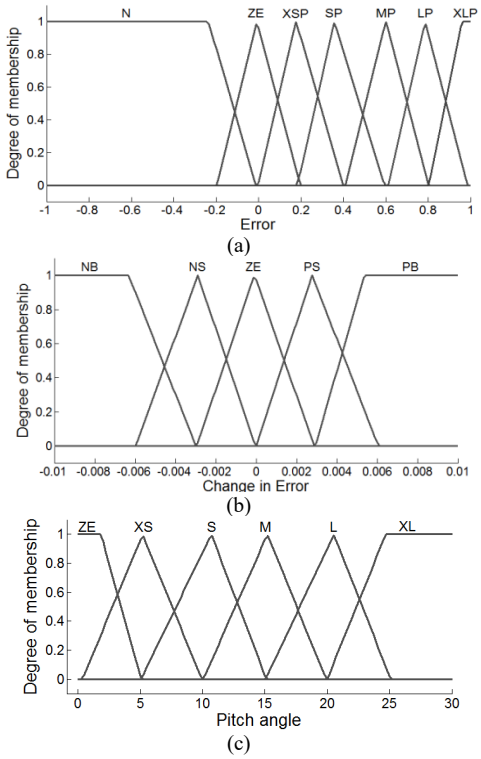


Figure. 12. The designed MFs of FLC (a) Error (b) Change in error (c) Pitch angle

B. Rule base

With help of the experts' knowledge on the pitch-angle control of the IG wind energy system, control strategies are framed as a set of IF-THEN rules and are as indicated:

If (ε is x_1) and ($c\varepsilon$ is y_1) then (β_c is w_1)

Similarly, 35 rules have been defined for all input-output MFs as shown in Table 1.

Table.1. Pitch angle controller rules

Change in Error ($\mathcal{CE}$)	Error ($\mathcal{E}$)						
	N	ZE	XSP	SP	MP	LP	XLP
NB	ZE	ZE	ZE	XS	S	M	L
NS	ZE	ZE	XS	S	M	L	XL
ZE	ZE	ZE	XS	S	M	L	XL
PS	ZE	ZE	XS	S	M	L	XL
PB	ZE	ZE	S	M	L	XL	XL

C. Defuzzification

The final stage is defuzzification of the combined input signal .The common defuzzification methods used for the FLC are (a) the first (or last) of maxima,(b) centroid-of-area and (c) mean-of-max methods. In this study, the centroid-of-area method has been used which is the most popular method from among the others. The defuzzification method converts the output fuzzy to crisp output value.

The constants used in the fuzzy logic controllers are scaling gains for input and output gains. The values considered for them chosen by trial and error method as $K_e = 1$ $K_d = 55$ and $K_u = 7.2$.

4. Results and Discussions

The employed wind energy system single-line diagram is as shown in Fig. 1. The pitch angle controller is shown in Fig. 11 which can serve for both, excessive wind and network fault conditions. In order to investigate the effectiveness of the control strategy two cases has been considered as below.

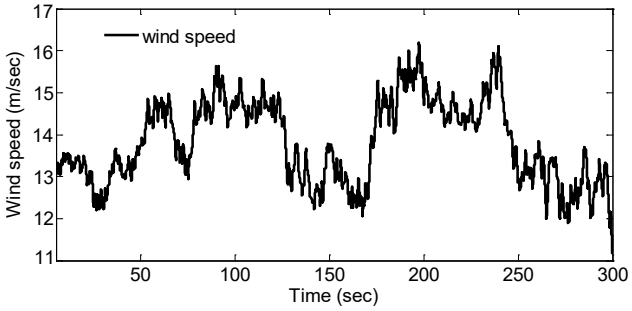


Figure. 13. Wind speed characteristics

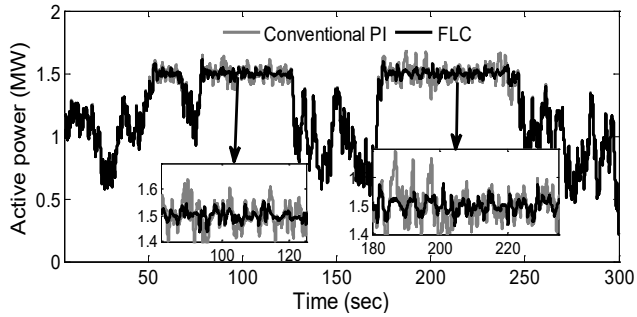


Figure.14. Generator active power

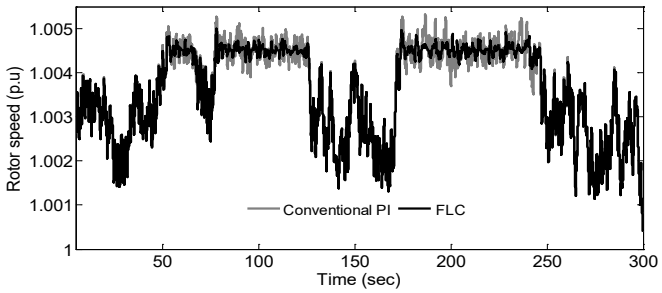


Figure.15. Generator rotor speed

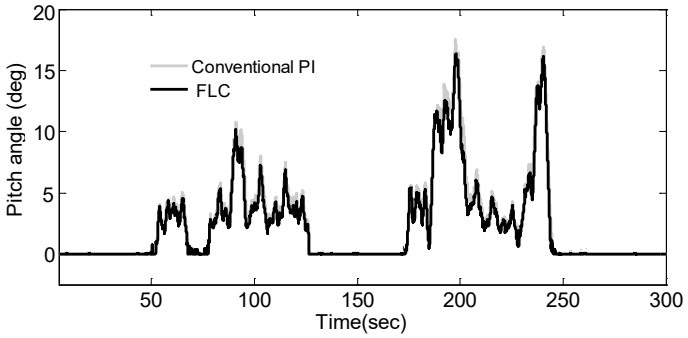


Figure.16. Pitch angle

4.1. Case A: Wind speed disturbances

The wind speed pattern employed for the study is as shown in Fig. 13. Using this wind speed characteristics, the performance of the proposed FLC based pitch angle control was investigated and compared with the results obtained with a conventional PI controller.

The responses of the generator active power, rotor speed and pitch angle controller are shown in Fig.14 -16. Fig. 14 shows the output power of the generator with PI and FLC controllers. Both the controllers can serve well in order to maintain the output power at rated level subjected to above rated wind speed (14 m/s). However, the FLC based controller has maintained the output power at rated level more effectively than PI controller since, the FLC work well with lower overshoot compared to PI controller. Similarly, rotor speed responses with both controllers are obtained and shown in Fig. 15. The pitch angle profiles of the WES with employed controllers are as shown Fig.11.

4.2. Case B: Transient Faults

The transient analysis of the WES has performed using the wind power system configuration shown in Fig. 1. This type of investigation is very important under the severe faults to guarantee the wind farm connection to the grid as per the recent grid codes. We demonstrate the effectiveness of the controller using the most severe fault i.e. a symmetrical fault .At the rated wind speed a three phase short circuit fault has been created time at t=10s for the duration of 150ms. Figures 17 -19 show the performance of the WES with PI and FLC controller.

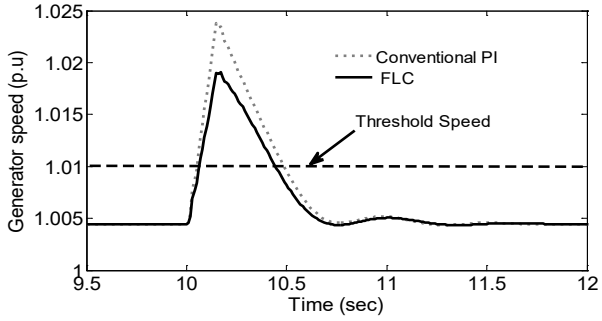


Figure.17. Generator rotor speed

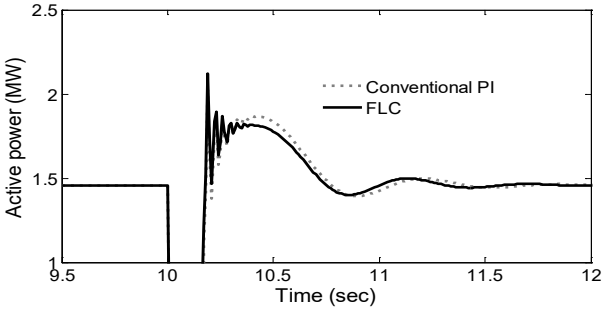


Figure.18. Generator active power

Fig. 17 shows the generator rotor speed variation. It starts increasing rapidly after the fault has been initiated. If the rotor speed exceeds the threshold rotor speed (ω_r^{TH}) then the proposed rotor speed-pitch angle controller comes into action and the

generator speed becomes stabilized with additional control using PI or FLC controllers. FLC shows better performance than the PI controller. The generator active power is shown in Fig. 18.

The pitch angle controller of the WT with PI and FLC is as shown in Fig. 19. It is noticed that for a large increase in pitch angle control the FLC controller decreases the mechanical torque and hence the rotor speed settles to a stable value.

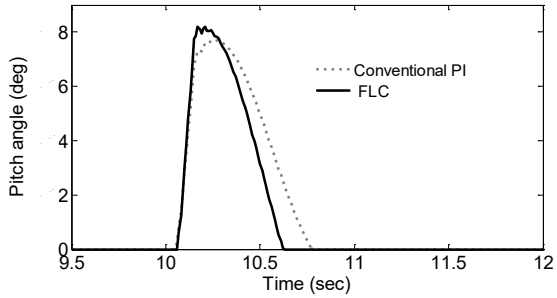
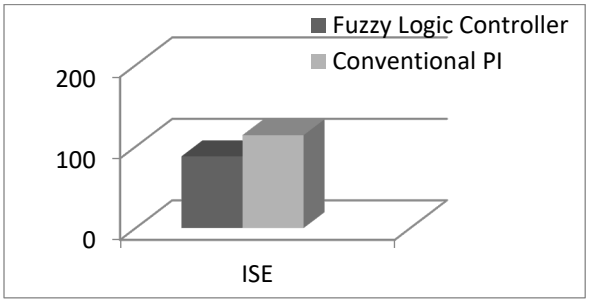


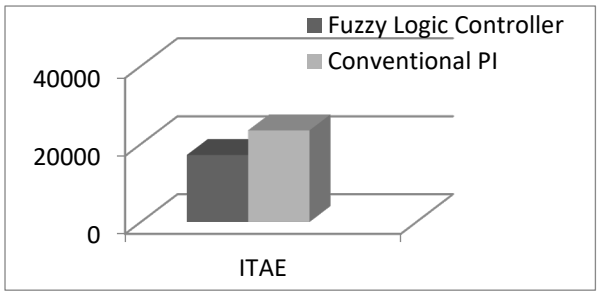
Figure. 19. Pitch angle

Table 2. Comparison of ISE and ITAE for different controllers

Method	ISE	ITAE
Conventional PI	113.90	23500
Fuzzy logic controller	**88**	**17200**



(a)



(b)

Figure. 20. Graphical representation of comparison of controller indices (a) ISE (b) ITAE

4.3. Controllers performance indices

Various controller performance indexes are available to estimate the controller performs improvement or degradation of the system such as Mean Relative Error (MRE), Integral Absolute Error (IAE), Integral Time Absolute Error (ITAE), Integral Square Error (ISE) and Integral Time Square Error (ITSE). To justify the performance of the controllers properly, the most popularly used indices ISE and ITAE are adopted in this paper and defined as

$$ISE = \int_0^T e^2(t)dt \tag{19}$$

$$IATE = \int_0^T t\left|e(t)\right|dt \tag{20}$$

Smaller the value of ISE and ITAE, better the controller performance. The quantitative results are presented in Table 2. It is observed from the Table 2 that the proposed FLC controller has minimum values of ISE and IATE compared to PI controller, therefore, the proposed controller exhibits better controlling performance. Moreover, the graphical representations of ISE and ITAE with both controllers are as shown in Fig. 20 (a) and (b), respectively.

5. Conclusion

This paper, proposes a fuzzy logic based novel pitch angle controller. It not only improves the quality of output power when the wind speed is above rated and high frequent but, can also improve the transient performance of the WES under faults. Two cases has been considered (wind speed disturbance and fault on the network) to demonstrate the effectiveness of the proposed controller. To show the superiority of the proposed controller, the results were compared with proportional-integral based pitch angle controller. It was found from the obtained results that the proposed fuzzy based novel control strategy exhibits superior response than PI based pitch angle controller.

Appendix-I

Table.3. SCIG generator parameters

Parameters	Values
P_{rated} , V_{rated}	1.5 MW, 0.69 kV
R_s, R_r	0.004843 p.u., 0.004377 p.u.
L_s, L_r	0.1248 p.u., 0.1791p.u.
L_m, H	6.77 p.u., 5.04 s
C	200 kVAR

Table.4.Wind turbine parameters

Parameters	Values
Rated power	1.5 MW
Rotor diameter	64m
Number of blades	3

Cut-in wind speed (V_{wCI})	4 m/s
Cut-out wind speed (V_R)	25 m/s
Rated wind speed (V_{wR})	14 m/s
Generator	SCIG

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. A. Naik and C. P. Gupta, "Transient stability enhancement of wind energy system using fuzzy logic based pitch angle controller," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 78-83 (2017). https://doi: 10.1109/UPCON.2017.8251026.

[2] The Global Wind Energy Council Belgium. http://www.gwec.net/global-Fig.s/wind-in-numbers/, (accessed 05.04.2017).

[3] O. Samuelsson, and S. Lindahl, "On speed stability", Power Syst. IEEE Trans., 20: 1179-1180 (2005).

[4] M. B. Kadri and S. Khan, "Fuzzy adaptive pitch controller of a wind turbine," 15th Int. Multitopic Conf. *(INMIC)*, Islamabad, 105-110 (2012).

[5] Y. Mi, X. Bao, Y. Yang, H. Zhang and P. Wang, "The sliding mode pitch angle controller design for squirrel-cage induction generator wind power generation system," Proceed. of the 33rd Chinese Control Conf., Nanjing, 8113-8117 (2014).

[6] Y. Mi, X. Bao, E. Jiang, W. Deng, J. Li, L. Ren, and P. Wang, "The pitch angle control of squirrel-cage induction generator wind power generation system using sliding mode control," 16th European Conf. on Power Electronics and Appli., Lappeenranta, 1-10 (2014).

[7] M. Al-Saffar and P. Musilek, "Fuzzy logic controller for large, grid-integrated wind farm under variable wind speeds," 17th International Scientific Conf. on Electric Power Engineering (EPE), Prague, 1-6 (2016).

[8] D. C. Vega, J. A. Marin and R. T. Sanchez, "Pitch angle controllers design for a horizontal axis wind turbine," IEEE Int. Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, 1-6 (2015).

[9] R. Sakamoto, T. Senjyo, T. Kaneko, N. Urasaki, T. Takagi and S. Sugimoto, "Output power levelling of wind turbine generator by pitch angle control using H-∞ control," IEEE Pow. Syst. Conf. Expo., 1-6 (2006).

[10] J. Tamura, "Transient stability simulation of power system including wind generator by PSCAD/EMTDC," IEEE Porto Power Tech Proceedings, vol.4, EMT-108, (2001).

[11] O. Wasynczuk, D.T. Man, J.P. Sullivan, "Dynamic behavior of a class of wind turbine generator during random wind fluctuations," IEEE Trans. on Power Apparatus and Systems, 100: 2873-2845 (1981).

[12] S. M. Muyeen, M. Hasan Ali, R. Takahashi, T. Murata, J. Tamura, Y.Tomaki, A. Sakahara, and E. Sasano, "Transient Stability Analysis of Grid Connected Wind Turbine Generator System Considering Multi-Mass Shaft Modeling," Electric Power Components And Systems, 34: 1121-1138 (2006).

[13] P. M. Anderson and A. Bose, "Stability simulation of wind turbine systems," IEEE Trans. Power Apparatus and Systems, 102: 3791-3795 (1983).

[14] T. Ackerman, "Wind power in power system", John Wiley & Sons. Ltd, (2005).

[15] R. Strzelecki and G. Benysek, "Power Electronics in Smart Electrical Energy Networks", Springer, London, UK, (2008).

[16] B. Wu, Y. Lang, N. Zargari and S. Kouro, "Power conversion and control of wind energy systems", John Wiley & Sons. Ltd, (2011).

[17] D. J. Trudnowski, A. Gentile, J. M. Khan, and E. M. Petritz, "Fixed-speed wind-generator and wind-park modeling for transient stability studies," Power Syst. IEEE Trans. 19: 1911-1917 (2004).

[18] A. P. Grilo, A. Mota, L. T. M. Mota and W. Freitas, "An analytical method for analysis of large-disturbance stability of induction generators", Power Syst. IEEE Trans. 22: 1861-1869 (2007).

[19] L. A. Zadeh, "Fuzzy sets", Information and control, 8: 338-353 (1965).

[20] S. Z. Hassan, H. Li, T. Kamal, M. Q. Abbas, M. A. Khan and G. M. Mufti, "An intelligent pitch angle control of wind turbine," 2017 International Symposium on Recent Advances in Electrical Engineering (RAEE), Islamabad, 2017, pp. 1-6. doi: 10.1109/RAEE.2017.8246144

[21] S. Z. Hassan, H. Li, T. Kamal, M. Nadarajah and F. Mehmood, "Fuzzy embedded MPPT modeling and control of PV system in a hybrid power system," *2016 International Conference on Emerging Technologies (ICET)*, Islamabad, 2016, pp. 1-6. doi: 10.1109/ICET.2016.7813236

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 33-44 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Direct Torque Control of Multiphase Doubly Converter-fed Asynchronous Machines Incorporating the Harmonic Torques

Javier Martínez-Román*[1], Luis Serrano-Iribarnegaray[1], Juan Carlos del Pino-López [2], Pedro Cruz-Romero[2]

[1]*Department of Electrical Engineering, Universitat Politècnica de València, Camino de Vera s/n, 46022, Valencia,Spain*

[2]*Department of Electrical Engineering, Universidad de Sevilla, Camino de los Descubrimientos s/n, 41092 Sevilla, Spain*

A R T I C L E I N F O

Article history:
Received: 05 June, 2018
Accepted: 22 June, 2018
Online: 11 July, 2018

Keywords:
Harmonic torques
Asynchronous machines
Space phasors
Direct Torque Control
Multiphase machines
GAFTOC principle

A B S T R A C T

Doubly fed asynchronous machines have an outstanding property: they can be operated up to twice rated speed delivering full rated torque. This paper presents, for the first time in the literature, a control system for multiphase asynchronous machines fed by Voltage Source Inverters (VSIs) both in stator and rotor that incorporates the harmonic torques. The system has three main and distinctive features: the independent control of the fundamental and harmonic torques, a very fast dynamic response for each one of these torques and a powerful method for selecting the best suited inverter state to achieve the evolution of the fundamental and harmonics flux linkage space phasors prescribed by the external control loops. The first feature is achieved through the decoupling of the multiphase machine provided by the Space Phasor Theory (SPhTh). The second one comes from the application of the General Approach for a very Fast TOrque Control (GAFTOC) principle. The third feature relies on using for multi-phase VSIs a simple but powerful switching-table based mode of operation that overcomes the limitations of the switching-table based modes of operation developed up to now, that only enable for the inverter to feed machines with no harmonic torques contribution.

1. Introduction

Multiphase machines are receiving increasing attention in the last two decades. This is due to several important advantages over their three-phase counterparts [1] like greater fault tolerance, reduced ratings of the electronic converters components, and the ability to make use of certain field harmonics (the greater the phase number, the higher the number of harmonics) in order to increase the total machine torque. This last possibility is only feasible for machines with relatively large winding factors for the low order harmonics (concentrated or nearly concentrated windings). The multiphase machines control development has been mainly directed so far to Induction Machines (IM) and to Permanent Magnet Synchronous Machines (PMSM). Multiphase Doubly Converter-Fed Asynchronous Machines (DCFAM) have not been dealt with in the technical literature so far. Yet, these machines show the great advantage over IM and PMSM of having an extended operating range to twice the rated speed while keeping the ability to provide full rated torque, thus effectively doubling the machine rated power. This outstanding property has only been considered so far in the case of three-phase DCFAM [2-4].

To extend the already developed Direct Torque Control (DTC) solutions for three-phase DCFAM to multiphase DCFAM with harmonic torques, Fig. 1, one must solve first three main problems:

a) decoupling the fundamental and harmonic torques for their independent control,

b) achieving a very fast dynamic response for each one of these torques, and

c) developing a switching table based operation strategy for multiphase VSIs feeding machines with harmonic torques contribution.

The first problem is easily addressed though the multiphase machine decomposition and decoupling carried out in the SPTh, as presented in [5].

Once the muti-phase machine decoupling has been achieved, solving the second problem boils down to simply apply to each one of the independent machines the GAFTOC principle [6], *in which the stator and rotor flux linkages space phasors, Ψ, play the fundamental role*. In this regard, it should be underlined that, from a conceptual viewpoint, Field Oriented Control (FOC) and DTC, the two well-stablished high performance electric drives control

*Corresponding Author: Javier Martinez-Roman, Email: jmroman@die.upv.es

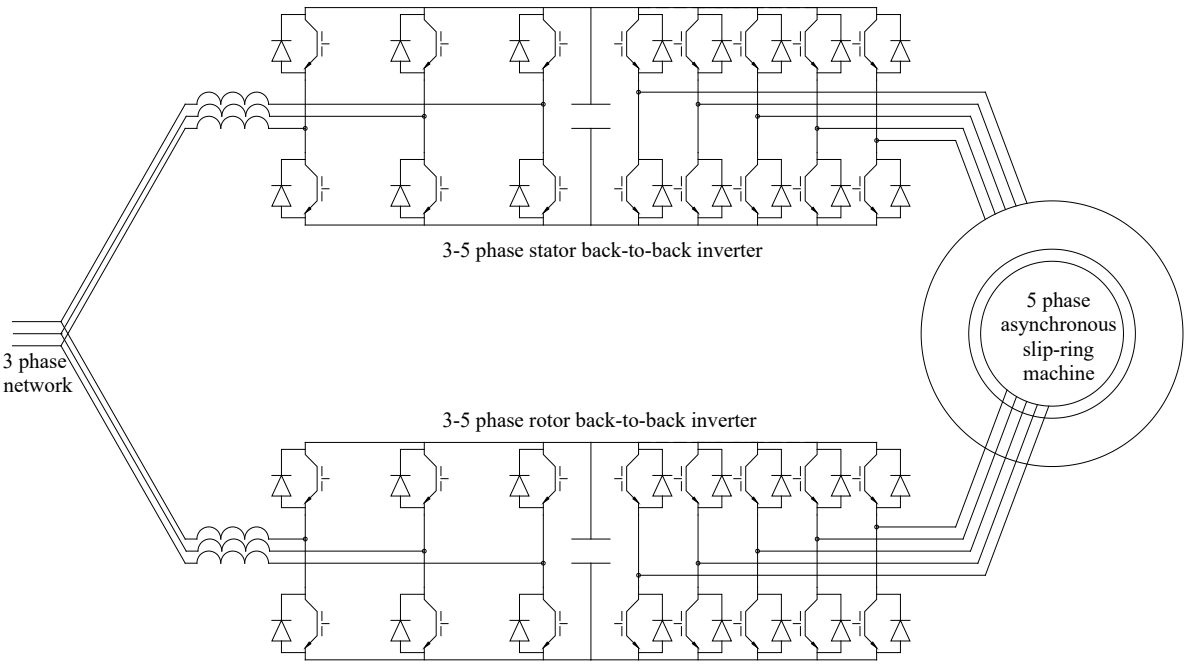


Figure 1. Doubly-converter-fed 5 phase asynchronous machine with 3-5 phase stator and rotor back-to-back inverters

methods, are two equivalent variants of the same and general GAFTOC principle, as explained in detail in section 7 of [6].

In this regard, for multiphase IM control, both FOC and DTC methods have been used. In the first case control schemes that allowed torque increases by incorporating harmonic torques were early developed [7]. Regarding the DTC method (which is the field of main interest of this paper) it has been also applied to multiphase IM, especially in the last years [8–10]. Yet in all of these cases, in clear contrast to some FOC implementations, there has never been torque improvement by space harmonics. As to muti-phase PMSM, they are, from the mathematical viewpoint and the control complexity, much simpler (no rotor windings) than IM. However, even in these cases, the DTC of the machine, to the best of the authors' knowledge, has also always been developed (with the exception of the results claimed in [11]), making use only of the fundamental torque [12-16]. As to the DTC of multiphase DCFAM, it has not been dealt with yet in the technical literature, not even in the simpler case of only fundamental wave operation.

With regard to the third problem, that is, with regard to the multiphase VSI operation, the first attempts at using a Space Vector Pulse Width Modulator (SVPWM) [17], were only able to control the fundamental voltage space phasor. Later on, [18] introduced a solution for the multiphase SVPWM with separated d-q spaces that allows the independent control of the d1q1-d3q3 voltage space phasors, but only with several constraints between their arguments and frequencies. These drawbacks were over-come in [19]. Still different improvements to the SVPWM operation of multiphase VSIs have been developed [20, 21]. In the case of multilevel converters, the choice of available inverter switching states increases considerably. In this respect [22] introduced a new multilevel multiphase SVPWM algorithm and [23] a generic digital Very High speed integrated circuit hardware Description Language (VHDL) module to implement it. All these strategies for selecting the inverter switching state require complex and time-consuming calculations, imply the use of Pulse Width Modulator (PWM) signal generators and assume, explicitly or implicitly, that the inverter can be modelled as an ideal controllable voltage source (which is not, since it has only a restricted a discrete number of switching states). Moreover, making use of an inverter operated this way for the machine control (usually in the FOC version) requires coordinate transformations (and thus a previous accurate determination of the angle between the involved reference frames). A switching-table based inverter operation strategy, initially introduced for three phase inverters [24], overcomes these drawbacks. In the case of multi-phase VSIs, this strategy has been first developed for inverters feeding 5-phase machines [8, 12, 13, 25]) and dual-three-phase machines [14, 16]. In a later and valuable paper [26], the switching-table based inverter operation strategy was extended to multiphase VSIs with an arbitrary odd number of phases, resorting to the so called virtual vectors definition methodology. In all of these cases, an ideal sinusoidal winding distribution has been assumed. Obviously, under this ideal assumption, only the fundamental space waves produce torque. Therefore, all of the methods developed until now are oriented to generate only fundamental voltage phasor precluding at the same time the presence of harmonic voltage phasors at the inverter output (suppression harmonics schemes). Indeed, these low order harmonic voltage phasors would drive large stator harmonic currents since the impedance (stator resistance plus leakage reactance) associated to these currents, which do not produce torque, is very small.

However, it seems neither logical nor the best option to disregard in all cases the possibility and advantages of *torque enhancement through harmonics contribution in the case of the DTC of multiphase machines*. In fact, these advantages have often been exploited in the case of FOC.

In view of the three aforementioned problems, this paper presents a switching-table based inverter operation strategy capable of simultaneously adjusting the fundamental and harmonics voltage space phasors, thus overcoming the limitations of the switching-table based modes of operation developed up to now (that only enable for the inverter to feed machines with no

harmonic torques contribution). Then, once this tool is available, the paper develops a DTC of a multiphase DCFAM, with voltage source inverters both in stator and rotor, Fig.1, with fundamental and also harmonic torques that, moreover, are controlled in an independent and very fast manner. This system keeps the main advantages of both three phase DCFAM (feasible operation up to twice the machine rated power and a balanced power share between the stator and rotor inverters) and multiphase configuration (greater fault tolerance, reduced ratings of the electronic converters components and torque enhancement through harmonics contribution).

The structure of this paper is as follows: first, section 2 describes by means of space phasors the fundamentals to split the actual multiphase machine into a set of much simpler machines, with only one effective flux linkage space phasor in stator, $\boldsymbol{\Psi}_{str}$, and another one in the rotor, $\boldsymbol{\Psi}_{rot}$ These fictitious machines are mechanically coupled, but electrically independent. Then, section 3 presents the new switching-table based operation strategy for multiphase VSIs feeding machines with useful harmonic torques. For the sake of simplicity, the strategy is explained on a five-phase machine. This strategy is computationally very light and can be easily extended to inverters with more than five phases and/or to multilevel inverters. Afterwards, section 4 develops in detail the scheme for the very fast and decoupled harmonic torques control which relies on the simultaneous and independent stator and rotor flux linkage space phasors control in each fictitious machine. The whole system has been checked in a five phase machine through an extensive set of simulations under different operating conditions, including step response to torque and speed references as well as speed reversals. Several of these simulations are displayed and commented in section 5. Finally, the main conclusions are summarized in section 6.

2. Multiphase Machines Decoupling Through the SPhTh

The approach overwhelmingly used for machines transient analysis is the so called magnetic coupling circuit approach (MCCA). Put it simply: the MCCA regards the machine as a network made up of resistances and inductances, many of which vary with the rotor position. This network is dealt with by means of complex matrix transformations.

The SPhTh for cylindrical or salient pole symmetrical machines with arbitrary number of phases and space harmonics has been presented in [5]. By contrast to the MCCA (machine as a network with rotor position dependent parameters), the SPhTh states that a machine can also be regarded as an electromechanical device that produces electromagnetic (field) waves with a restricted propagation capacity, namely they are forced to turn inside its air gap. These space waves, especially the scalar electric potential difference, the vector magnetic potential difference and the linear current sheet waves:

- can easily be characterized by dynamic space phasors: the voltage, flux linkage and current sheet space phasors, respectively
- can easily be correlated with the time phase quantities.

Notice that the SPhTh fully rejects space phasors to be mere mathematical entities without any physical meaning. In fact quite the opposite is true.

Another fundamental concept in the SPhTh is the dynamic m–phase system (DmPhS) of sequence *g*. By definition, the *m* quantities (*m* currents, *m* voltages, etc.) of an *m*-phase symmetrical winding are said to constitute a DmPhS of sequence "*g*", if they meet the following equations:

$$\begin{aligned} x_1(t) &= x(t)\cos[\varepsilon(t) - g\cdot 0\cdot\frac{2\pi}{m}] \\ x_2(t) &= x(t)\cos\left[\varepsilon(t) - g\cdot 1\cdot\frac{2\pi}{m}\right] \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_m(t) &= x(t)\cos\left[\varepsilon(t) - g\cdot(m-1)\cdot\frac{2\pi}{m}\right] \end{aligned} \tag{1}$$

where x(t) and ε(t) can be arbitrary time functions.

A DmPhS of sequence *g* only has two independent variables, *x(t)* and *ε(t)*. Thus, it can be fully characterized by a mathematical tool called the dynamic time phasor of the DmPhS of sequence *g:*

$$\left[\vec{X}\right]_g = \left[x(t)\ e^{j\,\boldsymbol{\varepsilon}(t)}\right]_g \tag{2}$$

Variables *x(t)* and *ε(t)* correspond with the instantaneous amplitude and position of the dynamic time phasor in its complex plane. Dynamic time phasors of the same sequence can be added up graphically and vectorially in an analogous manner as the classical Kapp's phasors do. If the *m* phases of the winding are numbered 1,2 ,….,*y*, ….. (*m*-1), *m*, the quantity of the general phase *y* of system (1) is obtained by simply projecting the phasor onto the phase axis in the complex plane of sequence *g*. That is ($\Re e$ stands for "real part of"):

$$x_y(t) = \Re e\left[x(t)e^{j\varepsilon(t)}e^{-jg(y-1)2\pi/m}\right] \tag{3}$$

Obviously, if amplitude and speed of the dynamic time phasor are constant, the DmPhS in (1) becomes a sinusoidal polyphase system of sequence *g*.

Sequences *g* and (*m* – *g*) are called complementary sequences. Any DmPhS of sequence *g* can always be converted into a DmPhS of sequence (*m* – *g*) in accordance with the equation (symbol * stands for conjugate complex):

$$\left[\vec{X}\right]_g = \left[\vec{X}^*\right]_{m-g} = \left[x(t)\ e^{-j\,\boldsymbol{\varepsilon}(t)}\right]_{m-g} \tag{4}$$

By definition, the complementary DmPhS of the one in (4) is:

$$\begin{aligned} &Complementary\ \left\{\left[\vec{X}\right]_g\right\} = \\ &= \left[\vec{X}\right]_{m-g} = \left[x(t)\ e^{j\,\boldsymbol{\varepsilon}(t)}\right]_{m-g} \end{aligned} \tag{5}$$

Comparing (4) and (5) it follows that the dynamic time phasors of two complementary DmPhSs are exactly equal except that their instantaneous arguments are opposite.

Let it be a seven-phase (*m*=7) stator symmetrical winding of a salient pole machine and consider the numerous stator electric potential difference space waves which appear at the stator surface with arbitrary changes in their amplitudes and speeds. It has been proven [5] that each of these waves with *p*, *8p*, *15p*, etc. pole pairs

(that is, each of the harmonic waves of relative order $h = 7q + 1$ with $q = 0, 1, 2, 3$, etc.) produces in the winding a voltage DmPhS of sequence 1. Likewise, each of the waves of order $h = 7q + 2$; $h = 7q + 3$, etc. produces a voltage DmPhS of sequence 2, 3, etc, respectively. Therefore, for a 7-phase stator winding all of the electric potential difference waves in axial direction at the stator surface can be classified into seven families. The instantaneous amplitudes and positions of the different waves belonging to the same family are, in the general case, quite different from one another. However, all of them originate DmPhSs of the same sequence which can be added up to give a resultant DmPhS of this same sequence. Thus, the combined action of all of the waves of one family on the phase voltages can be characterized by just one equivalent or effective voltage space phasor. Or alternatively: the resultant voltage DmPhS produced by a whole family of waves can be imagined to be produced by one only effective or equivalent space wave, the phasor of which is just the effective space phasor of the waves family.

Moreover, the two effective voltage DmPhSs produced by the families with $h = 7q + 1$ and $h = 7q + 6$ are DmPhSs of complementary sequence. Therefore, they can be added up so that just one space phasor suffices to characterize the combined action of both families (which form a group of space waves with $h = 7q \pm 1$). The same statement applies to the families with $h = 7q \pm 2$ as well as to the families with $h = 7q \pm 3$. Thus, the combined effect of all of the electric potential difference waves on the voltages of the seven phases in any dynamic state is fully characterized by just four effective voltage space phasors (one of which being the homopolar phasor). These conclusions apply too to the magnetic vector potential difference space waves (space phasors $\boldsymbol{\Psi}$), and can be extended to a symmetrical winding with an arbitrary number of phases, m.

Similarly, a current DmPhS of sequence g flowing through the stator symmetrical m-phase winding of a salient pole machine produces only current sheet space waves of relative order $h = mq \pm g$. Its complementary current DmPhS produces exactly the same space waves group, except that the instantaneous speed of any wave in this second case is opposite to that of the first case.

In summary, each one of the different $\boldsymbol{u}$, $\boldsymbol{\Psi}$ *and* $\boldsymbol{i}$ *effective dynamic space phasors existing in the multiphase machine is related to one specific and independent group of space harmonics*. This physical reality has been analyzed in detail in [5] and constitutes the true base that enables to "untangle" rather easily the structure of a constant air gap multiphase machine (either PMSM, IM or DCFAM) and to split it into an equivalent set of much simpler machines, *mechanically coupled but electrically independent*. Each one of these machines is associated to each one of the independent group of space harmonics mentioned above. As already pointed out in [5], it should be indicated here too that the authors in [27] had already dealt with the particular case of PMSM (no rotor winding and constant rotor magnetic field influence into the stator windings) following an approach completely different from the SPhTh. Resorting to different analysis tools (endomorphism, orthonormal base of eigenvectors, etc) they arrived at decoupling laws for PMSM mathematically fully equivalent to the general results presented in [5].

On the other hand, it is well known that only the interaction of stator and rotor space waves with the same pole number can produce torque. Assume again a symmetrical asynchronous machine with seven phases in stator and rotor. If one applies, exclusively, e.g., a current DmPhS of sequence $g = 5$ (or g=2) to the stator and another one to the rotor, the space waves (and their associated torques) of the two complementary families $h = 7q \pm 2$ can be controlled in any dynamic state *without modifying the space waves of the remaining families. This is just the physical base underlying the decoupled control of the multiphase machine*.

Moreover, by contrast to mains-connected machines, in multiphase converter-fed machines it is not necessary to take into account all the members of each wave family for determining the torque because in this case only the head members of their groups (that is, the harmonics with the lowest pole number) are able to provide a non-negligible instantaneous torque contribution. Notice, e.g. that in all control schemes of converter-fed three phase induction machines presented in the technical literature only the fundamental waves (that is, the head members of the stator and rotor waves families) are considered as to the torque production. The impact on the machine torque of the harmonic waves is considered negligible [30]. Therefore, in what follows, only the head members of each group of two complementary waves families will be considered (e.g., for a seven phase winding without homopolar components these head members are the fundamental, the third and the fifth harmonic space waves).

A three phase doubly converter-fed asynchronous machine without homopolar components only has one voltage and one flux linkage effective space phasor in the stator, and the same applies to the rotor. Its phasorial stator electric equation reads:

$$\overrightarrow{u_{str}} = R_{str}\overrightarrow{i_{str}} + \frac{d\overrightarrow{\Psi_{str}}}{dt} \tag{6}$$

In the DTC technique, since stator emf and voltage are almost equal quantities (if the resistive voltage drop is neglected), stator voltage space phasor guides the stator flux linkage phasor forcing it to approach a circular trajectory, so that this way only rotational emfs are induced, Figure 2.

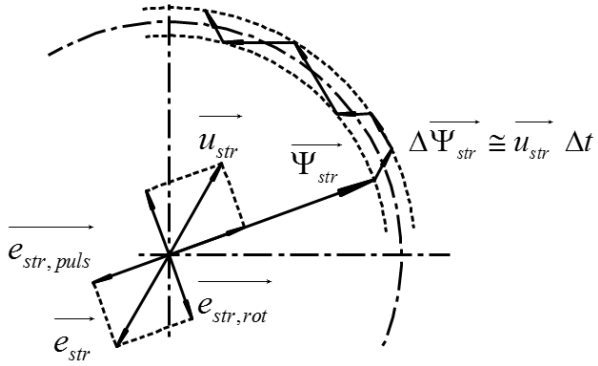


Figure 2. Stator flux linkage phasor changes in a modulation step and stator pulsational and rotational EMFs space phasors

Indeed, if $\overrightarrow{u_{str}} \cong -\overrightarrow{e_{str}}$, then according to (6):

$$\overrightarrow{u_{str}} \cong \frac{d\overrightarrow{\Psi_{str}}}{dt} \quad \Rightarrow \quad \overrightarrow{\Delta\Psi_{str}} \cong \overrightarrow{u_{str}} \cdot \Delta t \tag{7}$$

that is, during any converter control cycle the variation of the $\boldsymbol{\Psi}_{str}$ phasor takes place very closely in the direction of the voltage space phasor selected during the interval, Figure 2. In other words, the amplitude of $\boldsymbol{\Psi}_{str}$ is controlled by using radial voltage phasor components (that is, pulsational emf, which only affects the flux linkage module), whereas the torque is adjusted by applying tangential voltage components, that is, rotational emf, directly related to the torque generation, as explained in detail in [6], and which only affects the flux linkage phasor speed, producing a "tangential pull" on it.

Just as in three phase machines without homopolar components, the decoupled machines obtained from the actual multiphase machine also have only one effective voltage and one effective flux linkage space phasor in stator and rotor. Thus, as explained in [5], the electric polyphase winding (stator or rotor) equation of any of these independent machines reads:

$$\overrightarrow{u_{wind,x}} = R_{wind,x}\overrightarrow{i_{wind,x}} + \frac{d\overrightarrow{\Psi_{wind,x}}}{dt} \Rightarrow$$
$$\Rightarrow \overrightarrow{\Delta\Psi_{wind,x}} \cong \overrightarrow{u_{wind,x}} \cdot \Delta t \tag{8}$$

where x stands for stator or rotor. For more details on this subject, refer to [5].

3. Switching-Table Based Operation Strategy for Multiphase VSIs Feeding Machines with Useful Harmonic Torques

Single phase PWM presents a single degree of freedom: the ratio of t_{on} to t_{off}. Three phase PWM present three degrees of freedom, although only two of them are effective in terms of the load voltage space phasor due to the usual three wire connection. In DTC these two degrees of freedom result in the ability to obtain the most suitable load voltage space phasor out of the only 2^3=8 possible switching state combinations to modify the magnitude and angle of the stator flux linkage phasor prescribed by the external control loop. It is apparent that for every additional converter phase a new degree of freedom is available. Then the problem arises of finding a suitable connection between the control variables in a multiphase load and the corresponding VSI switching state selection.

In a 5-phase two level converter there are 2^5=32 switching states available (two of them resulting in the same null voltage space phasor). Therefore, they produce 31 different voltage space phasors of the first and of the third harmonic space waves. Let us consider a 5-phase machine, whose phase 1 is set at the real axis, fed by a converter with V_{bus} dc bus voltage. The status of each of the 5 inverter legs is defined by a set of five Boolean variables, $[sw]_x$, for which 1 means that phase x is connected to the dc bus positive terminal in Figure 1, and 0 that it is connected to the negative terminal. Each one of the 32 inverter switching states is then defined by the 5 bits of the following switching state index, sw:

$$sw = [sw]_5 \times 2^4 + [sw]_4 \times 2^3 + [sw]_3 \times 2^2 + [sw]_2 \times 2^1 + [sw]_1 \times 2^0 \tag{9}$$

Likewise, the voltage of any x-phase measured to the dc bus mid-point is given by

$$u_{x-phase,sw} = V_{bus}\left([sw]_x - \frac{1}{2}\right) \tag{10}$$

The fundamental and 3rd harmonic voltage space phasors are calculated from the phase terminal voltage according to:

$$\overrightarrow{u_{1,sw}} = \sum_{x=1}^{5} u_{phase-x,sw} e^{j(x-1)\frac{2\pi}{5}}$$
$$\overrightarrow{u_{3,sw}} = \sum_{x=1}^{5} u_{phase-x,sw} e^{j3(x-1)\frac{2\pi}{5}} \tag{11}$$

The 31 circles in Figure 3 mark the affixes of the 31 different voltage space phasors of the fundamental wave in its phasorial domain. Analogous information provide the 31 circles in Figure 4 with regard to the third harmonic. These voltage space phasors for the 32 switching states are then organized as column vectors:

$$\vec{u}_1 = \begin{bmatrix} \overrightarrow{u_{1,1}} \\ \dots \\ \overrightarrow{u_{1,32}} \end{bmatrix} ; \vec{u}_3 = \begin{bmatrix} \overrightarrow{u_{3,1}} \\ \dots \\ \overrightarrow{u_{3,32}} \end{bmatrix} \tag{12}$$

Notice that the index or number of each row in (12) expressed in binary system coincides with the switching state index, sw, that originates the voltage space phasor corresponding to the mentioned row.

The first step into the selection of the adequate switching state is to evaluate every state in terms of its ability to change magnitude and angle of the two flux linkage space phasors, $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_3$. To do so the complex phasorial plane of each space harmonic is divided into 20 equal sectors (The reason to divide each one of the two phasorial domains into 20 sectors is to assure that while each flux linkage space phasor remains within a given sector, the real and imaginary parts of all 30 non null voltage space phasors, calculated in a reference frame aligned with the flux linkage space phasor, do not change their signs). Then the real and imaginary parts of the voltage space phasors of the first and third harmonics corresponding to each switching state combination are calculated in a reference frame in which the real axis is directed along the middle of the sector, Figure 3 and Figure 4.

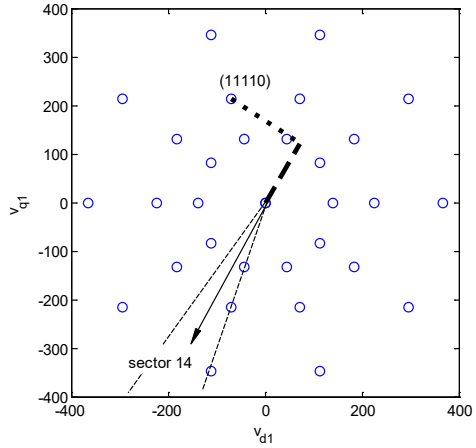


Figure 3. Real (thick dashed) and imaginary (thick dotted) parts of the fundamental *voltage* space phasor when the fundamental *flux linkage* space phasor lies in sector 14. The inverter switching state considered is 11110.

According to the winding equation (8) and section 2, Figure 2, *it is just these real and imaginary parts that quantify how well suited each switching combination is to quickly modify the corresponding flux linkage space phasor magnitude and angle respectively*. For example, the higher and positive the real part of the voltage phasor the better suited to increase flux linkage phasor magnitude, while the higher and negative the imaginary part the better suited to rotate the flux linkage space phasor backwards.

Let us consider the inverter switching state (11110) during a time interval at which the first harmonic flux linkage space phasor lies, e.g., within sector 14 while, at this same time, the third harmonic flux linkage space phasor lies in sector 11. Figure 3 shows for the mentioned switching state index, *sw* = (11110), the corresponding real and imaginary parts of the fundamental voltage space phasor (calculated in a reference frame aligned with the fundamental flux linkage space phasor). Analogous information provides Figure 4 related to the third harmonic voltage space phasor.

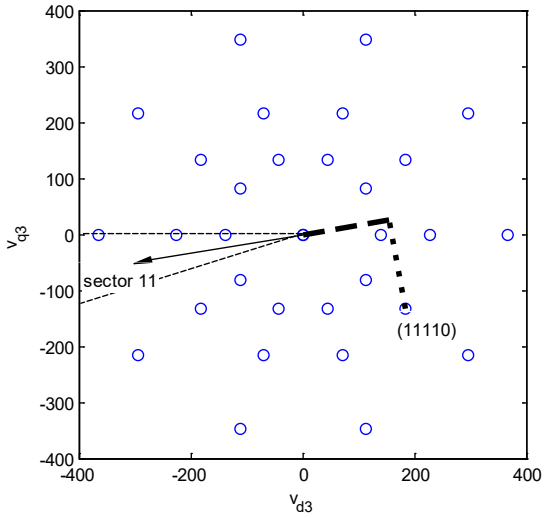


Figure 4. Similar to Fig. 3, but for the third harmonic space wave and when Ψ_3 lies in sector 11

The result of these operations performed for all sectors and all switching states are four 32x20 matrices (matrices M_{T1} M_{T3} $M_{\Psi 1}$ and $M_{\Psi 3}$ in Figure 6). These matrices, (13), are the imaginary or real parts of the matrices obtained as the complex product of the voltage space phasors column vectors times $(\mathbf{u}_{\gamma/2})^*$:

$$\begin{aligned}
[M_{T1}] &= \mathrm{Im}\left(\vec{u}_1 \times \left(\vec{u}_{\gamma/2}\right)^*\right) \\
[M_{\Psi 1}] &= \mathrm{Re}\left(\vec{u}_1 \times \left(\vec{u}_{\gamma/2}\right)^*\right) \\
[M_{T3}] &= \mathrm{Im}\left(\vec{u}_3 \times \left(\vec{u}_{\gamma/2}\right)^*\right) \\
[M_{\Psi 3}] &= \mathrm{Re}\left(\vec{u}_3 \times \left(\vec{u}_{\gamma/2}\right)^*\right)
\end{aligned} \tag{13}$$

where $(\mathbf{u}_{\gamma/2})^*$ is the row vector formed by the complex conjugates of the 20 sector central unitary vectors, according to the following equation:

$$\vec{u}_{\gamma/2} = \left[e^{+j\frac{2\pi}{40}} \quad e^{+j3\frac{2\pi}{40}} \quad e^{+j5\frac{2\pi}{40}} \quad \ldots \quad e^{+j39\frac{2\pi}{40}} \right] \tag{14}$$

In these matrices, each row is associated with one of the 32 available switching states in the 5-phase converter, while each one of the 20 columns is associated with one of the 20 sectors in which the phasorial domain for the fundamental and 3rd harmonic has been split into.

Notice that, once selected, e.g., sector 11, Figure 4 shows how the imaginary and real parts of the third harmonic voltage space phasor for one specific switching state can be obtained. Yet, this operation has to be performed, in this sector, for the 32 possible inverter switching states. The 32 x 2 values obtained this way are just the 32 values in column 11 of matrices M_{T3} and $M_{\Psi 3}$. Thus, matrices M_{T1} M_{T3} $M_{\Psi 1}$ *and* $M_{\Psi 3}$ express the ability of the 32 inverter switching state combinations to modify angle (torque) or magnitude (flux) of the fundamental or of the third harmonic flux linkage space phasors when these phasors lie in each one of the 20 sectors. From each one of these four matrices, and by means of the actual **Ψ** sector calculators, the system extracts the appropriate column, Figure 6, that is, the column corresponding to the sector in which the actual fundamental (or third harmonic) flux linkage space phasor lies (the actual sector is simply and directly determined from the **Ψ** phasor angle). Each one of these four columns is then weighted; or, more precisely, it is multiplied by the normalized desired trend (obtained from the *torque and flux linkage comparators*) in magnitude and angle of both flux linkage space phasors.

These weights for the stator and rotor fundamental and third harmonic torque and flux control are those shown in Figure 6 as *dt1, dp1, dt3, dp3*. They are equal to the actual torque or flux error (ΔT, $\Delta \Psi$) in per unit of the corresponding rated value multiplied by a suitable weighing factor, *k*, that is:

$$\begin{aligned}
dt1 &= \frac{\Delta T_1}{T_{1,rated}} k_{T1}; \quad dp1 = \frac{\Delta \Psi_1}{\Psi_{1,rated}} k_{p1} \\
dt3 &= \frac{\Delta T_3}{T_{3,rated}} k_{T3}; \quad dp3 = \frac{\Delta \Psi_3}{\Psi_{3,rated}} k_{p3}
\end{aligned} \tag{15}$$

where the k-values in (15) have been chosen in this paper equal to:

$$\begin{aligned}
k_{T1} &= \frac{T_{1,rated}}{T_{rated}}; \; k_{T3} = \frac{T_{3,rated}}{T_{rated}} \\
k_{p1} &= \frac{\Psi_{1,rated}}{\Psi_{rated}}; \; k_{p3} = \frac{\Psi_{3,rated}}{\Psi_{rated}}
\end{aligned} \tag{16}$$

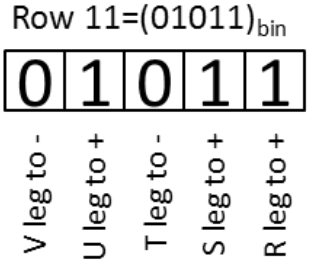


Figure 5. Five-phase Inverter switching legs status as determined by individual bits of row number conversion to binary: example for row number = 11. Inverter output phases are designed in this figure as *V, U, T, S* and *R*

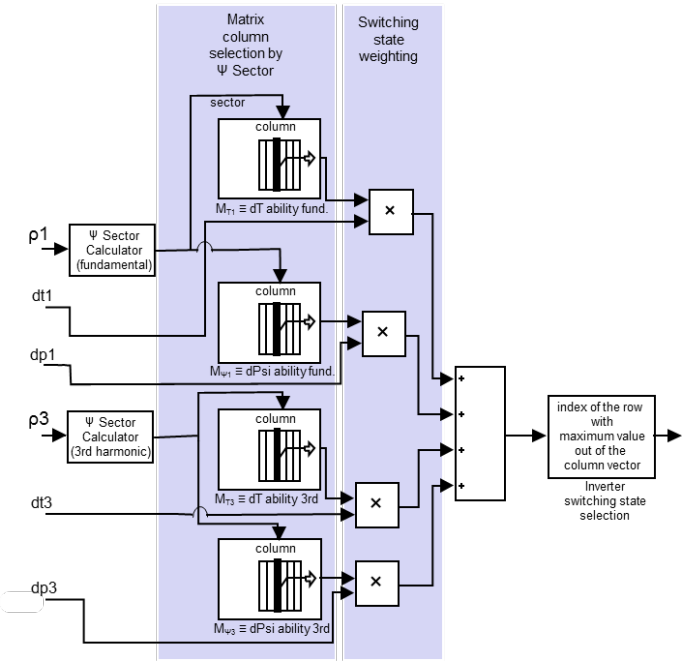


Figure 6. Operations flow for the 5-phase DTC: sector calculators, columns selection, weighing (products), combined switching state ability to modify magnitude/angle of the flux linkage space phasors (addition), best inverter switching state selection and, finally, binary conversion to get the inverter legs switching states. ρ_x: flux linkage space phasor angle (1: fundamental; 3: 3rd harmonic); dt_x and dp_x: normalized desired trend in magnitude and angle of the flux linkage space phasors.

The four column vectors mentioned above are added up to form a single column vector, and then its row with the highest value is selected. From this selected row the inverter switching state is obtained immediately: indeed, in accordance with the comments just after eq. (12), the row number conversion to binary directly yields the required switching state for all five inverter legs, Figure 5. This procedure overcomes the unsolved problem found in switching-table based operation strategy of multiphase VSIs in order to select the inverter switching state which is best suited to *simultaneously modify* magnitude and angle of the fundamental and also of the higher harmonic flux linkages space phasors in the way prescribed by the external control loops. (Remember that the space phasors **Ψ** play the central role in the GAFTOC strategy).

In summary, the switching state selection process is divided into two main steps. First, each switching state ability to force changes in the magnitude and angle of the flux linkages space phasors of fundamental and third harmonic is calculated (this step is carried out off-line and stored as an array in the controller memory). Second, depending on the magnitude and torque/angle errors for the flux linkage space phasors, the best suited out of the 32 available switching states is selected (this step needs only real products to be performed and is carried out on-line).

The whole process is then very light in terms of computational requirements as only basic operations and in small number are required. It needs to be carried out just once in every modulation period. The most computationally heavy operations for the matrices calculation are carried out off-line and only the resulting arrays are stored in the control system memory. It is important to notice too that this strategy can directly be extended to inverters with more than 5 phases and/or to multilevel inverters.

4. Independent and Very Fast Control of the Fundamental and Harmonics Torques of Multiphase Doubly Converter-Fed Asynchronous Machines

The Direct Torque and Stator and Rotor Flux Control (DTSRFC) of the three phase DCFAM in [4] actually set the basis for the very fast control of the fundamental and harmonics torques of the multiphase DCFAM.

The fundamental idea behind a very fast torque control strategy was already applied to develop the DTSRFC: In steady state of polyphase machines only rotational emfs appear in all of their windings and, therefore, these are just the emfs (that is, the ones related to the electromechanical energy conversion process) that have to be enhanced. The pulsational emfs (that is, the ones associated to amplitude variations of the flux linkage space phasors) should be kept null. This essential idea (pulsational emfs are a waste of resources and time as to the torque production) should be extended to transient states too if a torque control with very fast dynamic response and an efficient utilization of the machine (and of the power conditioning unit) is desired. This is the GAFTOC principle [6].

The three phase doubly converter-fed asynchronous machine is especially well suited to put into practice such control strategy. Indeed, to keep the pulsational emfs null in all of its windings, it suffices to operate the stator and rotor converters so as to maintain $\mathbf{\Psi_{str}}$ and $\mathbf{\Psi_{rot}}$ magnitudes constant. Should there still be any degree of freedom in the operation of the converters, it must be used so as to force the biggest possible rotational emfs (speed of the flux linkages space phasors).

In a three phase machine, the following equations hold [3,4]:

$$T = -\frac{m}{2} p \frac{L_\mu}{\left(L_{str} L_{rot} - L_\mu^2\right)} \left(\vec{\Psi}_{str} \times \vec{\Psi}_{rot}\right)$$

$$T = \frac{m}{2} p \frac{L_\mu}{\left(L_{str} L_{rot} - L_\mu^2\right)} \Psi_{str} \cdot \Psi_{rot} \cdot \sin\delta \tag{17}$$

It should be noticed [5] that space phasors are not physical vectors but (mathematically seen) complex time-varying quantities. Thus, the vectorial product in (17) should be interpreted as a formal operation.

Therefore, in the three phase machine the control is performed, [4], by keeping the $\mathbf{\Psi_{str}}$ and $\mathbf{\Psi_{rot}}$ modules constant and adjusting the torque through the angle, δ, between both phasors, Figure 7, according to (17). This same procedure must simply be extended to all of the fictitious independent machines of the multiphase machine.

A multiphase inverter enables, for each one of these fictitious machines, two degrees of freedom, that is, to freely adjust magnitude and angle of the corresponding flux linkage space phasors, **Ψ** [5]. Therefore, in a multiphase DCFAM (two inverters) this control strategy still leaves one degree of freedom to completely define the $\mathbf{\Psi_{str}}$ and $\mathbf{\Psi_{rot}}$ references in each fictitious machine. In the particular case of three phase machine this remaining degree of freedom is used to force a balanced share of electric power between both inverters. For more details, see [4].

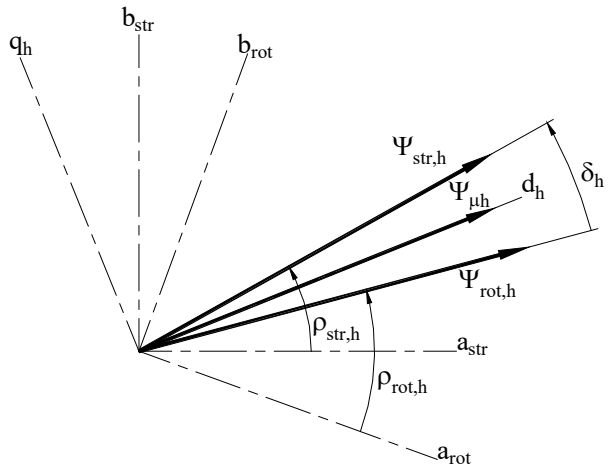


Figure 7. Stator, rotor and magnetizing flux linkage space phasors and flux load angle δ for the space harmonic of relative order h. Systems (a_{rot},b_{rot}), (a_{str}, b_{str}), (d_h, q_h): rotor, stator and magnetizing phasorial reference frames for the harmonic h.

In multiphase machines a new additional and important control task arises, namely how to adjust the suitable value and orientation of the harmonic magnetizing flux linkage space waves with regard to the fundamental one. In this sense, in a machine with the classical sinusoidal air-gap induction distribution only in the narrow region where the induction is maximum becomes the iron saturated to the desired value at any one instant. Yet, it seems clear that the magnetic circuit would be better utilized if the air-gap induction wave approximates a nearly rectangular one. Notice, nevertheless, that the control scheme in Figure 8 below enables to easily obtain any desired induction waveform. To this end, it suffices to choose the suitable amplitudes of the different harmonic flux linkage space waves (block flux linkage reference distribution in Figure 8) and their orientations with respect to the fundamental magnetizing flux linkage wave (18). These values are chosen to obtain (in machines with a high number of phases) an air gap induction wave of quasi-rectangular shape.

Notice too, from another perspective, that the possibility of obtaining almost any desirable air gap induction waveform in machines with a high number of phases relies on the fact that, with increasing phase number, the converters can inject additional current DmPhSs of different sequences, g, with the appropriate $\varepsilon(t)$ values in (1), which is valid for steady and transient states.

In this regard, it should be remembered from section 2 that the only condition to have different waves groups is to inject (through the suitable currents) dynamic time current phasors of different (and not complementary) sequences. For instance, in the particular case of steady state of a five phase machine, in order to get the two wave groups of $h = 5q \pm 1$ and h = $5q \pm 2$, it is only required to feed the winding, simultaneously, with two suitable sinusoidal polyphase current systems, A and B. In system A the phase angle between the currents of two consecutive winding phases should be γ or 4γ (that is, g_A = 1 or 4), with γ=2π/5. In system B it should be 2γ or 3γ (g_B = 2 or 3). Obviously, the functions $\varepsilon(t)$ are in this case of sinusoidal currents: $\varepsilon_A(t) = \omega_A t$ and $\varepsilon_B(t) = \omega_B t$, where ω_A and ω_B are two arbitrary constant values. If *in addition* it is desired that the head members of both wave groups turn in the same direction and at the same speed, then one must choose g_A =1, g_B =3 (or g_A =4, g_B =2) and ω_B=$3\omega_A$. In other words, and contrary to some statements in the literature, the frequency $\omega_B = 3\omega_A$ does not imply the existence of a third harmonic wave (which is tied to the sequence g of the current DmPhS) but only determines its speed [5].

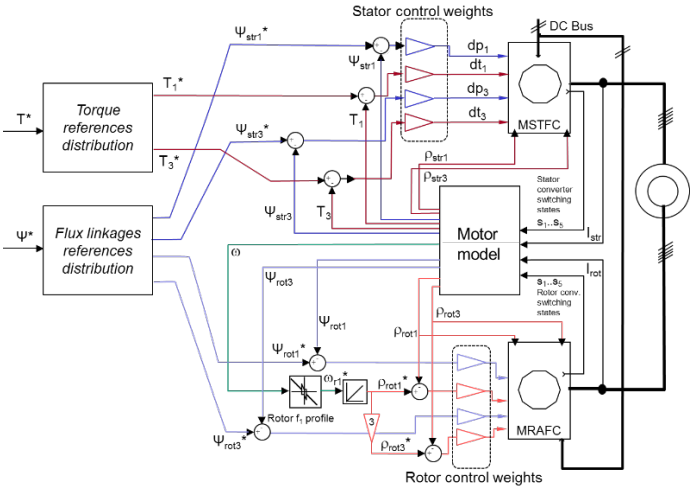


Figure 8. DTC diagram of a five phase doubly converter-fed asynchronous machine incorporating the harmonic torques. Symbol * stands for reference values. For a detailed explanation, see text.

The fact that, as above indicated, the air gap induction wave approximates a quasi-rectangular shape leads to a resultant magnetizing flux linkage wave that approximates a symmetrical triangular distribution so that one peak of all the magnetizing flux linkage space waves for the space harmonics of relative order 1, 3, … must occur at the same air gap angular position. In other words, the magnetizing flux linkage space phasors must be oriented in such a way that their angles (measured in their respective phasorial domains) meet the following equation:

$$\rho_{\mu 1} = \frac{1}{3}\rho_{\mu 3} = \frac{1}{5}\rho_{\mu 5} = \ldots \tag{18}$$

The diagram in Figure 8 shows the global control structure. It starts with the machine total torque and flux linkages references: the machine torque reference is the output of the speed controller (not shown in the figure) while the flux linkages reference can be either kept constant or suitably adjusted in the case of field weakening operation. The torque references distribution block provides the torque share for the fundamental and the third harmonic. An analogous statement hold for the flux linkages references distribution block. Then, the stator flux linkage and torques references are compared with the actual machine values and the errors are divided by the corresponding rated values and multiplied by the suitable weighing factors. The results of these operations, (15), are the inputs *dt1, dp1, dt3, dp3* to the stator Multiphase Stator Torque and Flux Controller (MSTFC) block, which has the structure of Figure 6. As for the rotor, the flux linkages comparators are similar to those of the stator. However, the role of the stator torque comparators is played here by the flux linkage space phasor angle comparators: the machine speed sets (by means of the rotor frequency profile, Figure 9) the rotor fundamental frequency. This rotor frequency is integrated to provide, first, the rotor fundamental flux linkage space phasor angle and thereafter, making use of (19), Figure 8 bottom, the angle of the rotor third harmonic space phasor. Then, the references for amplitude and

angle of the first and third harmonic rotor flux linkage phasors are compared with the actual values in the machine, and the errors are multiplied by the corresponding rotor control weights. Finally, the results of these operations are the inputs *dt1, dp1, dt3, dp3* to the rotor Multiphase Rotor Angle and Flux Controller (MRAFC) block, which also has the structure of Figure 6.

To synthetize and highlight the hard core of the diagram in Figure 8 let us consider as a first step only the fundamental waves. The input data required to operate the rotor converter are, in summary, the magnitude and angle errors of the rotor flux linkage space phasor. In turn, the input data required to operate the stator converter are the magnitude and angle errors of the stator flux linkage space phasor (this last one, that is, the error of the stator flux linkage space phasor orientation with respect to the rotor flux linkage phasor is indirectly given by the required fundamental torque, according to (17)). This way, both fundamental flux linkage space phasors can be forced upon the stator and the rotor by the corresponding converter. This scheme is repeated for all of the harmonic waves in a decoupled way (for the harmonics, the rotor phasor angle is given by specifying the position of the harmonic magnetizing space wave with respect to the fundamental one). And since, moreover, the modules of all of the harmonic flux linkage phasors are always kept constant, there are no pulsational emfs induced in the machine (GAFTOC principle) and, therefore, the torque contribution of the fundamental and the harmonics can be adjusted independently and very quickly.

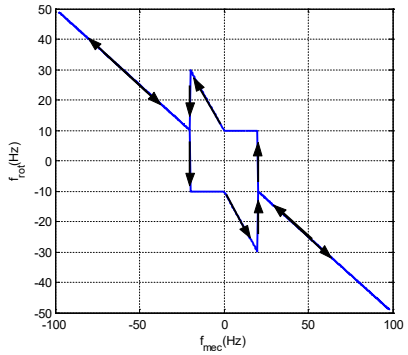


Figure 9. Rotor frequency profile

Let us introduce a final and brief comment on the rotor frequency vs. speed profile chosen, Figure 8 and Figure 9. This profile, as explained in [4], assures a balanced power share between the stator and the rotor converters and, at the same time, avoids operation at voltages close to converter ceiling voltage (higher frequencies) but maintaining a similar minimum converter frequency both in stator and rotor all over the speed range. A minimum frequency is important to allow accurate estimation of stator and rotor total flux linkage space phasors with simplified machine models.

5. Simulation Results and Discussion

There are several proposals to distribute the torque and flux linkages between the fundamental and the higher harmonic waves in asynchronous multiphase machines [28,29]. All of them have in common that the increase in machine torque is achieved through an increase of the fundamental air gap induction, Figure 10. This way, the machine pole flux is increased too. However, as the maximum value of the air gap magnetic induction is kept equal to that of sinusoidal operation, iron losses are accepted in the mentioned proposals to be scarcely affected. In accordance with these ideas, in this paper the third harmonic induction amplitude has been set to 0.18 of the fundamental wave. This results in a third harmonic torque contribution equal to 3.6% that of the fundamental for a total rms current equal to the one for sinusoidal operation.

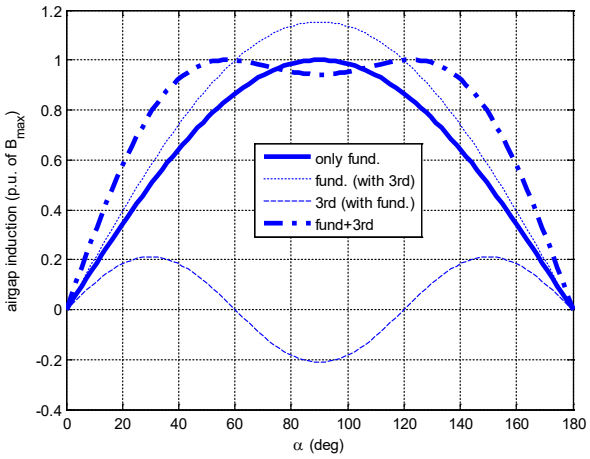


Figure 10. Air gap induction wave in a 5-phase machine for sinusoidal (solid line) and quasi-square wave operation (dash-dor line). The maximum value of the air gap induction is equal in both cases.

An extensive set of simulations, the results of some of which are displayed below, have been carried out on a five phase machine using Matlab-Simulink. The main machine parameters are listed in table 1.

Table 1 5.5 kW Drive Main Parameters

General	P=5,5kW, n=1438 rpm, η=0.863, cos φ=0.83 J=0.05 kg·m^2, stator and rotor phases: 5 Converter maximum frequency 8 kHz
Stator	U=138 V, I=10.9 A, f=50 Hz R=0.75Ω, L_σ=4.3mH (leakage,), $L_{\mu,1}$=80.4mH (main ind.) ; concentrated winding (one full pitch coil per pole pair).
Rotor	U=138 V, I=9.4 A, f=50 Hz R=0.54Ω, L_σ=3.1mH (leakage) $L_{\mu,1}$=80.4mH (main ind.); concentrated winding (one full pitch coil per pole pair).

The first point to note, Figure 11, is the ability of the control system to comply with one of its main objectives: the fast and decoupled control of the torques produced by the harmonics 1 and 3, while keeping constant the stator and rotor flux linkages space phasors magnitudes of both fundamental and third harmonic, as required by the GAFTOC. In the first row of graphs the step torque changes attest the very high dynamic response to torque reference changes while in the second and third row the excellent decoupling between torque and flux linkage, both for stator and rotor and for the fundamental and third harmonic, is clearly apparent. The last row of graphs shows on the one hand (by comparison with the first row) the direct relationship between torque and q-axis current and, on the other hand, the approximate sharing of the magnetizing current between stator and rotor. (Although only the stator d-axis current is plotted, it can be seen how it is approximately kept to

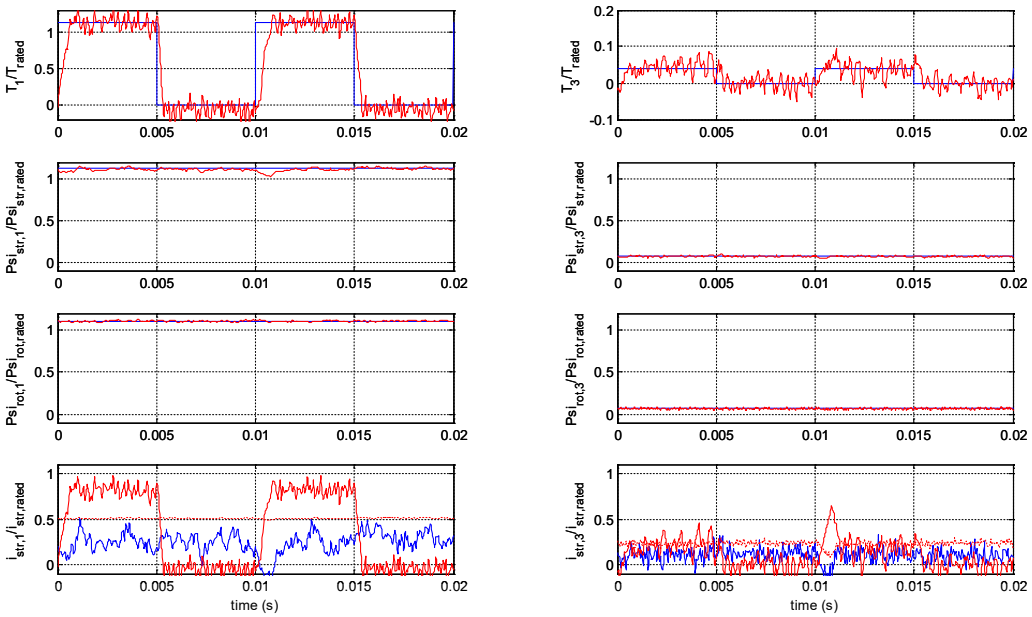
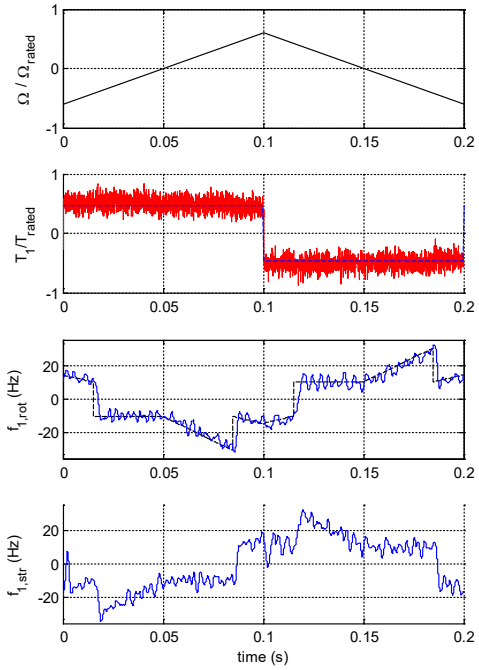


Figure 11. Rated torque step response at rated speed: left fundamental, right third harmonic. Up: Torque reference and actual value. Center: stator (above) and rotor (below) total flux linkage magnitude. Bottom: stator d, q currents and machine magnetizing current (dotted).

one half of the machine magnetizing current, meaning that the rotor provides the remaining half).



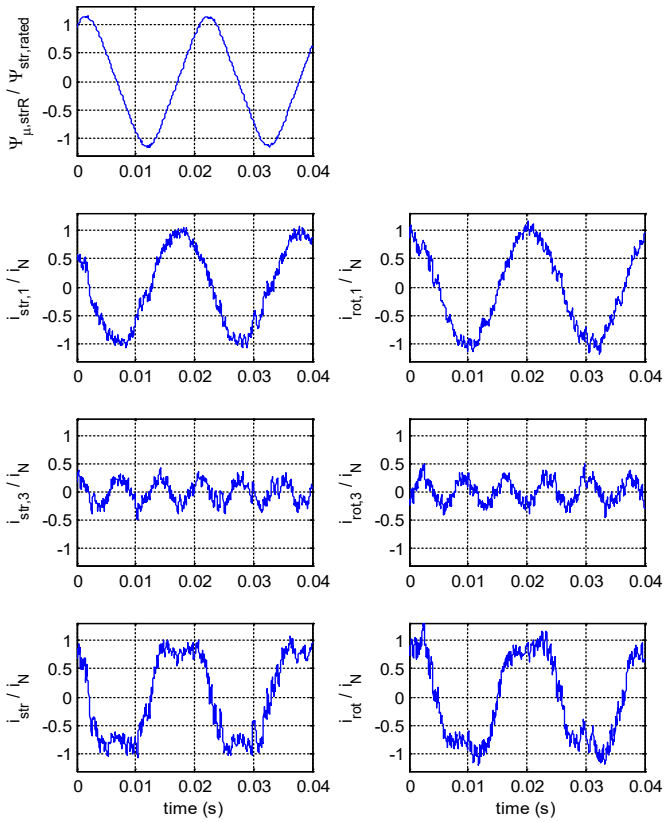


Figure 13. Machine operation at full load and twice rated speed. Left: Normalized magnetizing flux linkage of stator phase R (top), stator phase currents for sequence 1 (above) and 3 (center) and total phase current (below). Right: rotor phase currents for sequence 1 (above) and 3 (center) and total phase current (below).

Figure 13 shows the stator and rotor phase currents at full load (rated stator rms current) and twice rated speed operation and the quasi-triangular magnetizing flux linkage of phase R. This figure shows, for stator and rotor, the current share between the

fundamental and the third harmonic to provide the total required machine torque. The quasi-square waveform of the total phase current (as opposed to the sinusoidal current waveforms only possible up to now with previous multiphase DTC) clearly shows the third harmonic contribution to the torque enhancement achieved not exceeding the rms current rated value. It is worth noticing how a total torque of 116% of rated torque requires a relatively modest third harmonic torque of about 4% of rated torque to be added to the 112% of rated torque provided by the fundamental space waves.

The torque increase because of the third harmonic contribution appears also clearly when the 5-phase drive is compared with an equivalent three-phase drive with the same rated power and phase current as in Fig. 14 and Fig. 15. The right side in both figures shows the performance of the conventional 3-phase DCFAM under DTSRFC (as introduced in [4] and also explained in [6]) with the expected sinusoidal phase flux linkage and phase currents both in stator and in rotor. On the contrary, the left side in Fig. 15 clearly shows the quasi-triangular phase flux linkage and the associated quasi square-wave stator and rotor phase currents as in Figure 13. The reduced maximum current for the 5-phase drive (due to the harmonic content with equal rated rms total current) can be clearly appreciated also in comparison with the sinusoidal phase current in the 3-phase drive.

One can easily check that the results on the right side of Fig. 14 and Fig. 15 are a particular case of the results on the left side. Mathematically: reducing the degrees of freedom in the DTC of the multiphase DCFAM while keeping the GAFTOC strategy results in the DTC scheme of the three-phase machine (no contribution of the harmonic torques, associated to the fictitious harmonic machines). Three-phase DCFAM have already been dealt with in the literature, since they show the great advantage over three-phase IM and PMSM of having an extended operating range of up to twice the rated speed while keeping the ability to provide full rated torque and a balanced power share between the stator and rotor inverters. Yet, multiphase DCFAM have not been dealt with so far, not even in the simpler case of only fundamental wave operation. This problem has been addressed in this paper in the general case (arbitrary odd number of phases and contribution of harmonic torques). And, as shown by the figures and discussion in this section, the system presented keeps the main advantages of both the three-phase DCFAM (doubling the power) and of the multiphase configuration (greater fault tolerance, reduced ratings of the electronic converters components and torque enhancement through harmonics contribution).

6. Conclusions

Constant air gap multiphase PMSM, IM and DCFAM can be decomposed into an equivalent set of three-phase machines mechanically coupled but electrically independent. Thus, their torques can be controlled independently. On the other hand, each multiphase inverter enables for each one of these fictitious machines two degrees of freedom, so that magnitude and angle of either $\boldsymbol{\Psi}_{str}$ or $\boldsymbol{\Psi}_{rot}$ can be adjusted in each machine trough the inverter.

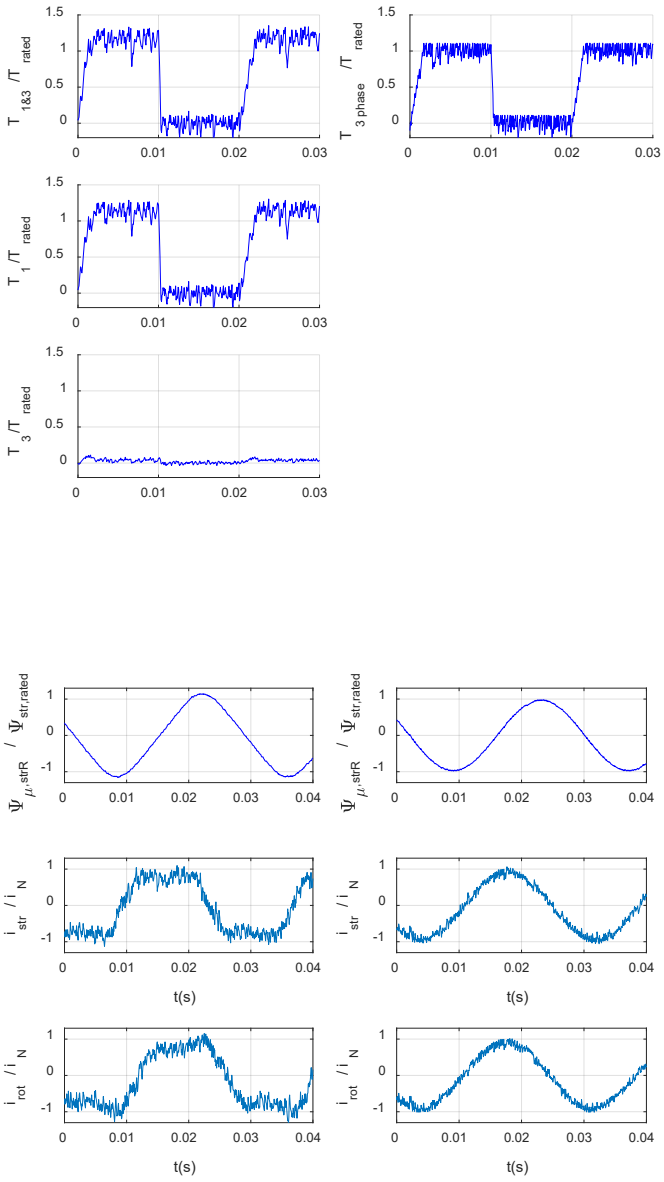


Figure 15. Direct torque and stator and rotor flux control of a DCFAM steady state operation at rated phase current: comparison for 5-phase (left) and 3-phase (right) drives. Top: stator phase flux linkage in p.u. of rated phase flux linkage. Middle and bottom: stator and rotor phase current, respectively, in p.u. of rated phase maximum current.

fictitious machines. GAFTOC states that the magnitudes of both $\boldsymbol{\Psi}_{str}$ and $\boldsymbol{\Psi}_{rot}$ of each machine must remain constant and that its torque must be adjusted by means of the angle between $\boldsymbol{\Psi}_{str}$ and $\boldsymbol{\Psi}_{rot}$, This strategy, with two inverters available, still leaves one degree of freedom to completely define the $\boldsymbol{\Psi}_{str}$ and $\boldsymbol{\Psi}_{rot}$ references. This degree of freedom has been used, in the main machine (which provides most of the power) to force a balanced power share between both inverters, and, in the remaining machines, to suitably align their magnetizing flux linkage space waves to the fundamental one.

In addition, in order to be able to apply the mentioned DTC to a DCFAM with an arbitrary odd number of phases, this paper also presents a new and very general switching-table based mode of operation for multiphase VSIs that, moreover, is computationally very light and can be easily extended to multilevel inverters.

Previous switching-table based mode of operation for multiphase VSIs enabled to take profit only of the machine fundamental wave (harmonic suppression schemes).

The whole system has been tested through an extensive set of simulations in a five phase DCFAM. They show that, as predicted by the theory, a very fast torque control of the DCFAM is possible: a) throughout the whole operational speed range b) keeping an excellent decoupling between torque and flux linkage, and c) independently for the fundamental and the harmonic torques. As to this point it should be underlined that the control of multiphase DCFAM is far more complete and versatile than that of multiphase IM and PMSM, simply because the degrees of freedom to design the control system are now double than before (two inverters instead of one). Reducing in different ways the degrees of freedom in the DTC of the DCFAM while keeping the GAFTOC strategy leads to the different DTC structures presented in the literature for the classic three-phase machines.

The main disadvantage of the DCFAM versus the alternative with IM or PMSM is the need for a slip rings rotor. However, even if the drive design around an IM or a PMSM were always the best option (which is dubious for high power applications), the multiphase machine decoupling and the fast and independent harmonic torque control proposed in this paper can still be applied to the DTC of multiphase IM and PMSM to great advantage.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E. Levi, R. Bojoi, F. Profumo, H. Toliyat, S. Williamson, "Multiphase induction motor drives - a technology status review," Electric Power Applications, IET, vol. 1, no. 4, pp. 489–516, July 2007.

[2] Y. Kawabata, E. Ejiogu, T. Kawabata, "Vector-controlled double-inverter-fed wound-rotor induction motor suitable for high-power drives," Industry Applications, IEEE Transactions on, vol. 35, no. 5, pp. 1058–1066, 1999.

[3] G. Poddar V. T. Ranganathan, "Direct torque and frequency control of double-inverter-fed slip-ring induction motor drive," Industrial Electronics, IEEE Transactions on, vol. 51, no. 6, pp. 1329–1337, Dec 2004.

[4] J. Martinez-Roman L. Serrano-Iribarnegaray, "Direct torque and stator and rotor flux control doubly converter fed induction machines," in IEEE Mediterranean Electrotechnical Conference, 2006. MELECON 2006. IEEE, 2006, pp. 1158–1161.

[5] L. Serrano-Iribarnegaray, "Space phasor theory and control of multiphase machines through their decoupling into equivalent three-phase machines," Electrical Engineering, vol. 96, no. 1, pp. 79–94, 2014.

[6] L. Serrano-Iribarnegaray J. Martinez-Roman, "A unified approach to the very fast torque control methods for dc and ac machines," Industrial Electronics, IEEE Transactions on, vol. 54, no. 4, pp. 2047–2056, Aug 2007.

[7] H. Xu, H. Toliyat, L. Petersen, "Five-phase induction motor drives with dsp-based control system," Power Electronics, IEEE Transactions on, vol. 17, no. 4, pp. 524–533, Jul 2002.

[8] L. Zheng, J. E. Fletcher, B. W. Williams, X. He, "A novel direct torque control scheme for a sensorless five-phase induction motor drive," IEEE Transactions on Industrial Electronics, vol. 58, no. 2, pp. 503–513, Feb 2011.

[9] D. Casadei, G. Serra, A. Tani, L. Zarri, "Direct torque control for induction machines: A technology status review," in Electrical Machines Design Control and Diagnosis (WEMDCD), 2013 IEEE Workshop on, March 2013, pp. 117–129.

[10] M. Bermúdez, H. Guzmán, I. González-Prieto, F. Barrero, M. J. Durán, X. Kestelyn, "Comparative study of dtc and rfoc methods for the open-phase fault operation of a 5-phase induction motor drive," in IECON 2015 - 41st Annual Conference of the IEEE Industrial Electronics Society, Nov 2015, pp. 002702–002707.

[11] X. Kestelyn, E. Semail, D. Loroil, "Direct torque control of multi-phase permanent magnet synchronous motor drive: application to a five-phase," in IEEE International Conference on Electric Machines and Drives, 2005., May 2005, pp. 137–143.

[12] Y. Gao L. Parsa, "Modified direct torque control of five-phase permanent magnet synchronous motor drives," in APEC 07 - Twenty-Second Annual IEEE Applied Power Electronics Conference and Exposition, Feb 2007, pp. 1428–1433.

[13] A. Yousefi-Talouki, S. A. Gholamian, M. Yousefi-Talouki, R. Ilka, A. Radan, "Harmonic elimination in switching table-based direct torque control of five-phase pmsm using matrix converter," in 2012 IEEE Symposium on Humanities, Science and Engineering Research, June 2012, pp. 777–782.

[14] K. D. Hoang, Z. Q. Zhu, M. Foster, "Optimum look-up table for reduction of current harmonics in direct torque controlled dual three-phase permanent magnet brushless ac machine drives," in 6th IET International Conference on Power Electronics, Machines and Drives (PEMD 2012), March 2012, pp. 1–6.

[15] X. Ma, Y. Yu, H. Zhang, W. Wang, L. Liu, "Study on direct torque control of dual y shift 30 degree six-phase pmsm," in 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA), June 2015, pp. 1964–1968.

[16] K. D. Hoang, Y. Ren, Z. Q. Zhu, M. Foster, "Modified switching-table strategy for reduction of current harmonics in direct torque controlled dual-three-phase permanent magnet synchronous machine drives," IET Electric Power Applications, vol. 9, no. 1, pp. 10–19, 2015.

[17] Y. Zhao T. Lipo, "Space vector pwm control of dual three-phase induction machine using vector space decomposition," Industry Applications, IEEE Transactions on, vol. 31, no. 5, pp. 1100–1109, Sep 1995.

[18] H.-M. Ryu, J.-H. Kim, S.-K. Sul, "Analysis of multiphase space vector pulse-width modulation based on multiple d-q spaces concept," Power Electronics, IEEE Transactions on, vol. 20, no. 6, pp. 1364–1371, Nov 2005.

[19] D. Casadei, D. Dujic, E. Levi, G. Serra, A. Tani, L. Zarri, "General modulation strategy for seven-phase inverters with independent control of multiple voltage space vectors," Industrial Electronics, IEEE Transactions on, vol. 55, no. 5, pp. 1921–1932, May 2008.

[20] O. Lopez, D. Dujic, M. Jones, F. Freijedo, J. Doval-Gandoy, E. Levi, "Multidimensional two-level multiphase space vector pwm algorithm and its comparison with multifrequency space vector pwm method," Industrial Electronics, IEEE Transactions on, vol. 58, no. 2, pp. 465–475, Feb 2011.

[21] G. Carrasco C. Silva, "Space vector pwm method for five-phase two-level vsi with minimum harmonic injection in the overmodulation region," Industrial Electronics, IEEE Transactions on, vol. 60, no. 5, pp. 2042–2053, May 2013.

[22] O. Lopez, J. Alvarez, J. Doval-Gandoy, F. Freijedo, "Multilevel multiphase space vector pwm algorithm," Industrial Electronics, IEEE Transactions on, vol. 55, no. 5, pp. 1933–1942, May 2008.

[23] J. Alvarez, O. Lopez, F. Freijedo, J. Doval-Gandoy, "Digital parameterizable vhdl module for multilevel multiphase space vector pwm," Industrial Electronics, IEEE Transactions on, vol. 58, no. 9, pp. 3946–3957, Sept 2011.

[24] I. Takahashi T. Noguchi, "A new quick-response and high-efficiency control strategy of an induction motor," Industry Applications, IEEE Transactions on, vol. IA-22, no. 5, pp. 820–827, Sept 1986.

[25] L. Gao, J. E. Fletcher, L. Zheng, "Low-speed control improvements for a two-level five-phase inverter-fed induction machine using classic direct torque control," IEEE Transactions on Industrial Electronics, vol. 58, no. 7, pp. 2744–2754, July 2011.

[26] R. Karampuri, J. Prieto, F. Barrero, S. Jain, "Extension of the dtc technique to multiphase induction motor drives using any odd number of phases," in 2014 IEEE Vehicle Power and Propulsion Conference (VPPC), Oct 2014, pp. 1–6.

[27] E. Semail, A. Bouscayrol, J.-P. Hautier, "Vectorial formalism for analysis and design of polyphase synchronous machines," The European Physical Journal Applied Physics, vol. 22, no. 3, pp. 207–220, 2003.

[28] D. Casadei, M. Mengoni, A. Tani, G. Serra, L. Zarri, "Torque maximization in high-torque density multiphase drives based on induction motors," in Energy Conversion Congress and Exposition (ECCE), 2010 IEEE, Sept 2010, pp. 3896–3902.

[29] L. A. Pereira, S. Haffner, L. F. Pereira, R. S. da Rosa, "Torque capability of high phase induction machines with sinusoidal and trapezoidal airgap field under steady state," in Industrial Electronics Society, IECON 2013-39th Annual Conference of the IEEE. IEEE, 2013, pp. 3183–3188.

[30] J.-A. Echeverria-Villar, J. Martinez-Roman, L. Serrano-Iribarnegaray, "Transient harmonic torques in induction machines: measurement and impact on motor performance," Electrical Engineering, vol. 94, no. 2, pp. 67–80, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 33-45 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Fuzzy Uncertainty Management in Multi-Shift Single-Vehicle Routing Problem

Francesco Nucci*

Department of Engineering for Innovation, University of Salento, Lecce, Italy

ARTICLEINFO

Article history:
Received: 14 August, 2018
Accepted: 19 October, 2018
Online: 01 November, 2018

Keywords:
Vehicle Routing Problem
Fuzzy Uncertainty
Scheduling

ABSTRACT

Our research deals with the single-vehicle routing problem (VRP) with multi-shift and fuzzy uncertainty. In such a problem, a company constantly uses one vehicle to serve demand over a scheduling period of different work shifts. Our issue relies on a routing problem in maintenance jobs, where a crew executes jobs in different sites. The crew runs during several work shifts but repeatedly returns to the depot before the shift ends. The goal is executing all the activities minimizing the makespan. We examine the impact of uncertainty in driving and maintenance processing time on system performance. We realize an Artificial Immune Heuristic to find optimal solutions considering both makespan and overtime avoidance. First, we introduce a framework to assess the uncertainty impact. Then, we produce a numerical company case study to examine the problem. Outcomes present significant improvements are obtained with the proposed approach.

1 Introduction

Vehicle routing problem (VRP) consists in determining a set of routes to visit a fixed set of customers, in order to minimize the total path length. Several versions of the VRP exist. If each customer specifies availability time windows, we deal with a VRP with Time Windows (VRPTW). Basic variants of the VRP consider the route planning for a vehicle fleet in a single period (shift). In this case, vehicles should come back to the depot before the shift ends. This problem is derived from a healthcare routing question. The healthcare company regularly dispatch products to medical sites. In this case, if overtime is allowed, performance could be significantly upgraded [1]– [4]. For instance, if a location scheduled in the next shift is on the current return path to the depot, a limited overtime allows vehicle serving it. This can significantly diminish next shift workload.

Investigation on the connection between health status and work hours documented concern about the influence of working long hours on people fitness [5]. Furthermore, working long hours increases the chance of micro-sleep in car drivers [6]. In these conditions, the company could be accused of vehicle collision due to enormous workload planning [7]. In fact, shift scheduling objective should limit both anxiety and harmful consequences on healthiness maintaining the work shift as steady as possible.

In an effort of limiting extra time, uncertainty effect on scheduling has to be restricted. Frequently, in optimization problems, data are supposed to be known with certainty. Nevertheless, in practice this is rare. Most commonly, the real data depend on uncertainty because of their irregular essence. Because the optimization problem solution usually reveals a great inclination to the data disturbances, neglecting the data ambiguity may conduct to suboptimal or infeasible solutions for a real case. Stochastic VRP (SVRP) was introduced in [8]. See [9] and [10] for a complete review of SVRP.

Robust optimization is a significant technique to deal with optimization problems subject to uncertainty [11]. The main motive for considering robust optimization is the non-stochastic nature of uncertain parameters. In such a case, a methodology is required to analyze compromise between performance and robustness.

Fuzzy set theory is a useful approach to handle uncertain information, while the stochastic approach is suitable to manage the stochastic parameters in VRP [12]. Some VRP papers utilize fuzzy sets theory for studying the influence of uncertain factors [13]–[23].

*Corresponding Author: F. Nucci, Via per Monteroni, Lecce, Italy, Tel.: +39.0832.297805, Email: francesco.nucci@unisalento.it

https://dx.doi.org/10.25046/aj030603

In this work, the VRP in a fuzzy random context is analyzed and the fuzzy random theory is adopted to manage such uncertain data. Fuzzy random variables illustrate a well-formalized notion involving data obtained from a random test in which data are supposed to be fuzzy sets.

We examine a multi-shift VRP with no overtime allowed considering fuzzy driving and job processing time. The question we considered is derived from a routing problem in maintenance activities. A maintenance team performs jobs in different sites using a vehicle for movements. A crew works in shifts and should come back to the depot before the shift ends. The goal is completing the maintenance activities in various places reducing the makespan. Since we deal with maintenance jobs, we ignore customer waiting times. We investigate the influence of the uncertainty of driving and job processing time on makespan and overtime avoidance.

The body of this paper is structured as follows. In Section 2, we report the problem formulation. In Section 3, we propose the Artificial Immune Heuristic to solve the problem. We present in Section 4 a case study analyzed with the proposed approach. We give concluding remarks in Section 5.

2 Problem formulation

2.1 Classical problem

Problem notation is reported in Table 1.

Let $N = \{1, 2, ..., n\}$ be the set of maintenance jobs to be executed at customer places. Parameter $d_{i,j}$ stands for the travel time between the location of jobs i and j. Parameter q_i expresses job i processing time. Parameters e_i and l_i represent job i time window.

Maximum shift duration is L, whereas p is the number of shifts in the scheduling horizon and $P = \{1, 2, ..., p\}$ is shift set. We generate $p+1$ depot copies described by nodes $n+1, \ldots, n+p+1$:

- node $n+1$ is the origin depot of shift 1,
- node $n+h$ represents the destination depot of shift $h-1$ together with the origin depot of shift h, with $h \in 2, \ldots, p$.
- node $n+p+1$ stands for the destination depot of last shift p.

The problem can be represented as a directed graph $G = (V, E)$, where $V = N \cup \{n+1, \ldots, n+p+1\}$. Arcs $(n+h, n+h+1)$ are used in the graph to illustrate cases in which a vehicle is not exploited during shift h.

Lastly, b_h and c_h represent the begin and end time of shift h, with $b_h = (h-1)L$ and $c_h = hL$, $h \in P$.

In this work, the next hypotheses are made. Driving times are positive, i.e. $d_{ij} > 0, \forall i, j \in V$. The triangular inequality is valid for d_{ij} , i.e. $d_{ij} \leq d_{ik} + d_{k,j}, \forall i, j, k \in V$. At least any single job can be executed in a shift because we suppose shift length satisfies $L \geq d_{0,i} + q_i + d_{i,n+1}, \forall i \in N$. Time windows are greater than the activity processing time: $l_i - e_i \geq q_i$, $\forall i \in N$. Consequently, a linear formulation for VRPTW is considered. Moreover, a feasible distance set is defined and all the jobs can be processed within their time window and shift duration.

In the following, decision variables are presented:

- $x_{i,j} = 1$ if the crew drives from node i to j, and 0 otherwise,
- α_i arrival time of the crew at node $i \in V, i \neq n+1$,
- δ_i departing time of the crew at node $i \in V, i \neq n+p+1$,
- σ_h real shift h duration,
- $y_h = 1$ if shift h is used by the crew, and 0 otherwise,

The problem is expressed as a mixed integer program (MIP) described in (1)-(13):

- The objective function (1) minimizes the system makespan.
- Constraints (2) and (3) provide that the depot nodes $\{n+1, \ldots, n+p+1\}$ are always visited. Depot node $n+h+1$ is visited after $n+h$. Depot nodes separate different shifts. All job nodes visited between node $n+h$ and $n+h+1$ are served during shift h. If the vehicle is not used in shift h, no job node is visited between node $n+h$ and $n+h+1$. Such constraints assure that any shift path starts from the origin depot and ends at the destination depot.
- Constraints (4) ensure that each customer is visited once, that is each node $i \in N$ is inspected once.
- Constraints (5) force that for all intermediate nodes (first node $n+1$ and last node $n+p+1$ are excluded) the inflow is equal to the outflow.
- Constraints (6) are the sub-tour elimination constraints and ensure coherence of time variables. Parameter M is an upper bound of $\delta_i + d_{ij}, \forall i, j \in V$.
- Constraints (7) represent the connection between arrival and departing times.
- Constraints (8)-(9) define time window constraints for the customer nodes, whilst (10)-(11) represent the time windows constraints for the depot nodes.
- Constraints (12) provide the real shift duration. Considering (10)-(12), parameter L is an upper bound of shift duration σ_h, $\forall h = 1, \ldots, p$.
- Constraints (13) determine whether shift h is exploited or not. In fact, if $x_{n+h,n+h+1} = 1$, then no demand node is inspected between depot h and $h+1$ (shift h).

Table 1: Problem notation

Symbol	Description
n	Number of Jobs
$N = \{1, 2, ..., n\}$	Job set
$i, j \in N$	Job index
$d_{i,j}$	Travel time between job i and job j locations
q_i	Job i processing time
e_i	Begin of job i time window
l_i	End of job i time window
L	Maximum shift duration
p	Number of shifts in the planning horizon
$P = \{1, 2, ..., p\}$	Shift set
$h \in P$	Shift set index
b_h	Begin time of shift h, equals to $(h-1)L$
c_h	End time of shift h, equals to hL
$x_{i,j} \in \{0, 1\}$	Crew travels from node i to j, if $x_{i,j} = 1$
δ_i	Departing time of crew at node $i \in V, i \neq n+p+1$
α_i	Arrival time of crew at node $i \in V, i \neq n+1$
σ_h	Actual shift h duration
$y_h \in \{0, 1\}$	Crew is active in shift h, if $y_h = 1$
M	Upper bound of $\delta_i + d_{ij}, \forall i, j \in V$.

$$\text{Minimize} \quad \alpha_{n+p+1} \tag{1}$$

subject to:

$$\sum_{j \in N \cup \{n+h+1\}} x_{n+h,j} = 1 \quad \forall h \in P \tag{2}$$

$$\sum_{i \in N \cup \{n+h\}} x_{i,n+h+1} = 1 \quad \forall h \in P \tag{3}$$

$$\sum_{j \in V \setminus \{n+p+1\}} x_{j,i} = 1 \quad \forall i \in N \tag{4}$$

$$\sum_{j \in V} x_{j,i} - \sum_{j \in V} x_{i,j} = 0 \quad \forall i \in V \setminus \{n+1, n+p+1\} \tag{5}$$

$$\delta_i + d_{ij} \leq \alpha_j + M x_{i,j} \quad \forall i, j \in V \tag{6}$$

$$\alpha_i + q_i \leq \delta_i \quad \forall i \in N \tag{7}$$

$$e_i \leq \alpha_i \quad \forall i \in N \tag{8}$$

$$\delta_i \leq l_i \quad \forall i \in N \tag{9}$$

$$b_h \leq \delta_{n+h} \quad \forall h \in P \tag{10}$$

$$\alpha_{n+h+1} \leq c_h \quad \forall h \in P \tag{11}$$

$$\alpha_{n+h+1} - \delta_{n+h} \leq \sigma_h \quad \forall h \in P \tag{12}$$

$$y_h = 1 - x_{n+h,n+h+1} \quad \forall h \in P \tag{13}$$

2.2 Problem with fuzzy uncertainty

We describe the uncertainty on driving times and job processing times with triangular fuzzy numbers, see Fig. 1. We consider the following notation:

- fuzzy travel time: $\widetilde{d_{ij}} = (d_{ij}^A, d_{ij}^B, d_{ij}^C)\ \forall i, j \in V$
- fuzzy job processing time: $\widetilde{q_i} = (q_i^A, q_i^B, q_i^C)\ \forall i \in N$

For this reason, decision variables related to arrival and departing time at node i turn into fuzzy variables $\widetilde{\alpha_i}$ and $\widetilde{\delta_i}$. Also, actual shift duration variables $\widetilde{\sigma_i}$ become fuzzy. Therefore, fuzzy numbers are introduced both in objective function (1) and constraints (6)-(12).

An important issue arises when models with fuzzy parameters are considered: defining the comparison method for objective function values [24]. In order to solve this question, we have to consider the ranking of fuzzy numbers [25, 26, 27]. The fuzzy ranking method is part of the solution approach in order to solve mathematical programs having coefficients of the objective function and coefficients of the constraints represented by fuzzy numbers.

The Expected Existence Measure (EEM) operator can be applied to a temporal fuzzy number $\widetilde{t}$, see (14) and [28, 29].

$$EEM_{\tilde{t}}(t) = \frac{\int_{-\infty}^{t} \mu_{\tilde{t}}(\tau) d\tau}{\int_{-\infty}^{+\infty} \mu_{\tilde{t}}(\tau) d\tau} = \frac{\int_{t_A}^{t} \mu_{\tilde{t}}(\tau) d\tau}{\int_{t_A}^{t_C} \mu_{\tilde{t}}(\tau) d\tau} \quad \in [0, 1] \tag{14}$$

Briefly, EEM defines the possibility with which a fuzzy event occurred at a certain time. Assigned a temporal value t_0, the EEM stands for the possibility a temporal fuzzy number $\widetilde{t}$ is lower than t_0. We denote this as $\Phi(\widetilde{t} \leq t_0) = EEM_{\widetilde{t}}(t_0) = \gamma_0$. Consequently, $\Phi(\widetilde{t} \geq t_0) = 1 - \gamma_0$. In Fig. 1, the membership function $\mu_{\widetilde{t}}(t)$ of the triangular fuzzy number $\tilde{t}$ and the corresponding $EEM_{\widetilde{t}}(t)$ are represented. Note that $\Phi(\widetilde{t} \leq t^A) = EEM_{\widetilde{t}}(t^A) = 0$ and $\Phi(\widetilde{t} \leq t^C) = EEM_{\widetilde{t}}(t^C) = 1$.

Various studies solve fuzzy mathematical problems by exploiting fuzzy ranking methods. First, fuzzy mathematical programming problems are converted into classical mathematical problems. Second, conventional techniques are applied. In the following section, we propose an innovative approach to consider the fuzzy variables in the *entire* solution process. We present the decision-maker with different optimal solutions based on 2-factor comparison: the objective function value and the possibility degree with which constraints are satisfied.

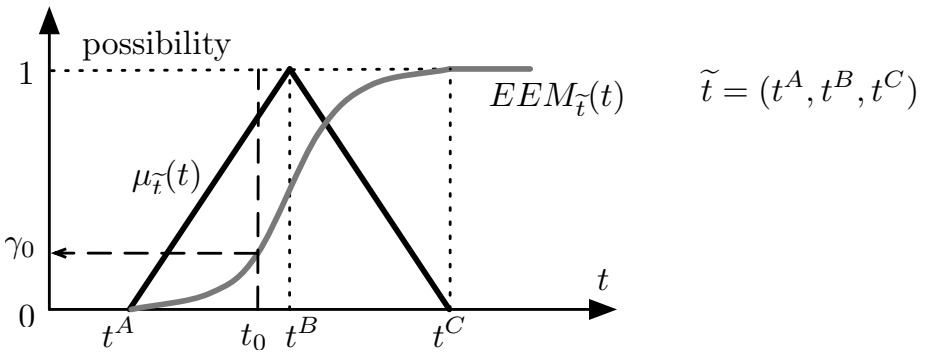


Figure 1: Triangular Fuzzy Number

In the previous work [30] we published on this question, we dealt with a different objective function: number of exploited shifts and maximum shift duration. In this research, we consider a different problem in which makespan is minimized. Furthermore, the solution algorithm has been improved with an appropriate solution 'affinity' definition in Section 3.2. Moreover, we extended the experimental campaign in order to investigate our approach performance. For such a reason a mathematical background is introduced in sections 4.2, 4.2.2, and 4.2.3.

3 Solution Method: Artificial Immune Heuristic

The animal immune system is a versatile pattern-recognition system that protects against foreign viruses and bacteria. The immune system is able to recognize and kill germs. The immune system cells, named *antibodies*, are casually diffused in the body. The system reacts to pathogens and improves the process of recognizing and eliminating pathogens by using two principles: clonal selection and affinity maturation. When a pathogen invades the organism, clonal selection creates a number of immune cells that recognize and eliminate the pathogen. While cellular reproduction occurs, the cells experience high rate physical mutations, as well as a selective process. Cells with superior affinity to the invading pathogen spread into memory cells.

Artificial Immune Algorithm (AIA) is a metaheuristic based on such system [31, 32, 33]. This paper intends proposing a fuzzy artificial immune algorithm to find optimal solutions for the aforementioned problem. AIA notation is reported in Table 2.

3.1 Solution Encoding

A solution is encoded as a string by using a fixed-length integer code and providing the order in which nodes are reached. Solution Ψ, representing a scheduling problem with $n = 5$ jobs on $p = 3$ available shifts, is reported in Fig. 2. Moreover, Fig. 2 shows the corresponding graph path, see section 2. The origin for shift 1 is node $n+1 = 6$ that is the path starting node. Then, the crew visits nodes 2, 3 and 5. Shift 1 ends at node $n+2 = 7$. After, shift 2 begins and crew inspects nodes 1 and 4. Shift 2 ends when node $n+3 = 8$ is reached. Finally, node $n+p+1 = 9$ is reached (path end) because no job is executed in shift 3.

Because node $(n+1)$ and $(n+p+1)$ are respectively the 'starting' and 'end' node, solution Ψ is just defined by the succession of the residual nodes of the graph, that is $v_\ell \in V \setminus \{n+1, n+p+1\}$, $\forall \ell = 1, \ldots, n+p-1$. Note that when $x_{ij} = 1$, $i, j \in V, i \neq n+1, j \neq n+p+1$, then nodes i and j are directly connected in the graph, consequently $\exists \ell = 1, \ldots, n+p-2$: $v_\ell = i$ and $v_{\ell+1} = j$, that is nodes i and j are consecutive in the string encoding. In Fig. 2 we have $x_{14} = 1$, consequently $\exists \ell = 5$, $v_5 = 1$ and $v_6 = 4$.

Considering fuzzy numbers, we have to appropriately evaluate both the solution objective function f and solution feasibility degree γ.

First, considering the solution Ψ and adopting the fuzzy addition operator described in (15), we calculate actual duration of shift 1, $\widetilde{\sigma_1}$, and shift 2, $\widetilde{\sigma_2}$, see (16)-(17). Because shift 3 is not used by the crew, we have the same arrival time at node 8 and 9: $\widetilde{\alpha_9} = \widetilde{\alpha_8}$. We calculate fuzzy makespan as $\widetilde{\alpha_9} = \widetilde{\alpha_8} = L + \widetilde{\sigma_2}$. Indeed, makespan is the crew arrival time at the depot at the end of shift 2 (node 8), that is the sum of shift 2 begin time L and actual shift 2 duration $\widetilde{\sigma_2}$. Since $\widetilde{\sigma_2}$ is a fuzzy number, the crisp objective function f in (18) can be adopted using the modal value σ_2^B of the fuzzy number $\widetilde{\sigma_2}$.

Second, we determine the degree with which the solution Ψ is feasible. So we calculate the possibility the shift duration constraints are enforced: $\Phi(\widetilde{\sigma_1} \leq L) = \gamma_1$ and $\Phi(\widetilde{\sigma_2} \leq L) = \gamma_2$. Time windows constraints are considered in the same way. Supposing for solution Ψ, fuzzy shift durations $\widetilde{\sigma_1}$ and $\widetilde{\sigma_2}$ are those reported in Fig. 3, it is possible that shift 1 actual duration is greater than maximum limit L. Indeed, half of fuzzy number $\widetilde{\sigma_1}$ stays on the right side of L. The value $\gamma_1 = EEM_{\widetilde{\sigma_1}}(L) = 0.5 < 1$ indicates the possibility that shift 1 ends before time L. Since, $\gamma_2 = EEM_{\widetilde{\sigma_2}}(L) = 1$,

Table 2: AIA notation

Symbol	Description
$popsize$	No. population antibodies
ng	No. generations
$pr1$	Rate of Rule1: full random node selection
$pr2$	Rate of Rule2: higher probability for closer node selection, $pr2 = 1 - pr1$
nc	No. clones in each generation
mr	Mutation Rate
nm	No. mutations in each generation
nea	No. exchangeable antibodies.
Ψ	Solution
v_ℓ	Solution encoding, $\forall \ell = 1,\dots,n+p-1$, $v_\ell \in V \setminus \{n+1, n+p+1\}$
f	Solution makespan
γ	Solution feasibility degree
$\mathcal{S}$	Solution set, $\lvert\mathcal{S}\rvert = popsize$
Γ	solution Pareto set, $\Gamma \subset \mathcal{S}$
$\mathcal{A}^{\mathcal{S}}(f,\gamma)$	Affinity for solution (f,γ) in set $\mathcal{S}$
exp	No. experiments in the campaign
$\mathcal{E}$	experiment set, $\mathcal{E} = \{1,\dots,exp\}$
ϵ	Experiment index, $\epsilon \in \mathcal{E}$
$\overline{\Gamma}^{\mathcal{E}}$	Hyper Pareto set: dominant solutions for set $\bigcup_{\epsilon\in\mathcal{E}} \Gamma_\epsilon$
ρ_ϵ	Impact of experiment ϵ

shift 2 duration constraint is certainly enforced. Finally, the solution Ψ feasibility degree γ in (19) is determined $\gamma = 0.5$.

$$\begin{aligned}
\widetilde{t_s} &= \widetilde{t_1} + \widetilde{t_2} = (t_1^A, t_1^B, t_1^C) + (t_2^A, t_2^B, t_2^C) \\
&= (t_1^A + t_2^A, t_1^B + t_2^B, t_1^C + t_2^C) && (15)\\
\widetilde{\sigma_1} &= \widetilde{d}_{6,2} + \widetilde{q}_2 + \widetilde{d}_{2,3} + \widetilde{q}_3 + \widetilde{d}_{3,5} + \widetilde{q}_5 + \widetilde{d}_{5,7} && (16)\\
\widetilde{\sigma_2} &= \widetilde{d}_{7,1} + \widetilde{q}_1 + \widetilde{d}_{1,4} + \widetilde{q}_4 + \widetilde{d}_{4,8} && (17)\\
f &= \alpha_{n+p+1}^B && (18)\\
&= L + \sigma_2^B \\
\gamma &= \min_{\substack{i\in N\\ h\in P}} \{\Phi(\widetilde{\alpha}_i \geq e_i), \Phi(\widetilde{\delta}_i \leq l_i), \Phi(\widetilde{\sigma}_h \leq L)\} && (19)\\
&= \min\{0.5, 1\} = 0.5
\end{aligned}$$

Eventually, for a given solution, we associate a crisp objective function value f and a feasibility degree γ. We perform a Pareto comparison on (f,γ) pairs for determining the solution Pareto set Γ: f should be minimized and γ should be maximized.

For example, we consider an alternative solution Ψ' in which all the available 3 shifts are exploited. The alternative solution Ψ' is reported in Fig. 4. Unlike the previous solution Ψ, in Ψ' Job 5 has moved from shift 1 to shift 3. On one hand, we have objective function $f = 2L + \sigma_3'^B$, see (18). On the other, for Ψ', we have γ is equal to 1 because all shift duration constraints are definitely enforced ($\gamma_i = 1$, $\forall i = 1,\dots,N$), see $\sigma_h^C < 1$, $\forall h = 1,2,3$ in Fig. 4. Comparing the solution Ψ' to the previous one Ψ, we note that f value gets worse ($f^{\Psi'} > f^{\Psi}$) and γ values is better ($\gamma^{\Psi'} > \gamma^{\Psi}$). Indeed, with the alternative solution, shifts are completed in time but one additional shift is required. Since we adopt a 2-factor comparison, both solution Ψ and solution Ψ' are Pareto optimal.

3.2 Affinity

Usually, in meta-heuristic approaches as AIA, an affinity level to be maximized is defined in order to take into account both solution objective value and constraint enforcement. Unfeasible solutions are penalized in terms of affinity. Solutions are ranked on the basis of their affinity values and, eventually, the one with the greatest affinity is defined as 'best solution'. In our approaches, two different factors (f and γ) are both considered in determining the affinity. Solutions having $\gamma = 0$ are certainly unfeasible.

We define the solution set as $\mathcal{S}$. Solutions in the Pareto set $\Gamma = \{(f,\gamma) | \gamma > 0\} \subset \mathcal{S}$ represent the optimal set, and we assign them the same maximum affinity value as in (20). Note that $\bar{f}$ is the makespan of the solution having feasibility degree 1. Since f stands for makespan to be minimized, whereas affinity is to be maximized, we put $\bar{f}$ in the second term as the makespan upper bound pL occupies the first term.

Using the normalized Manhattan distance from Pareto set (22), the affinity for solutions not in the Pareto set is computed (23). That is, considering a solution (f,γ) not in Γ, the minimum Manhattan distance from Pareto set is calculated and is used as a penalty for affinity. Quantity Δf is used as a weight in Manhattan distance for balancing the effects of makespan and feasibility degree.

3.3 Artificial Immune Algorithm

The proposed AIA is described in the following:

Step 1 - Initialization:

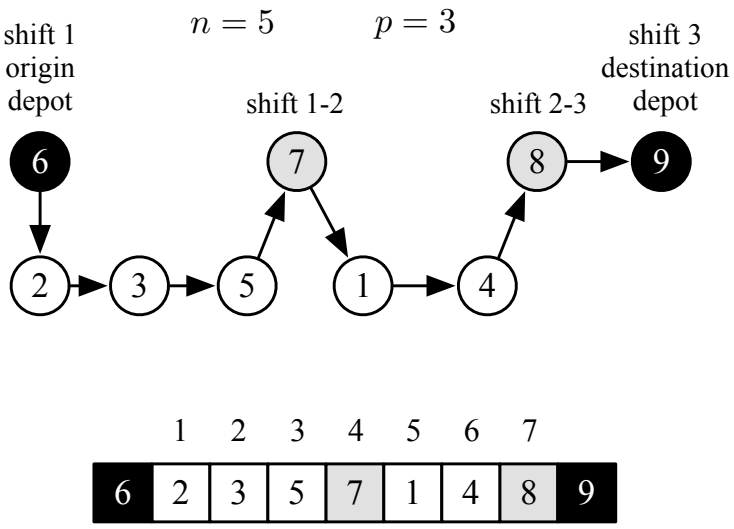


Figure 2: Solution Ψ graph path and encoding

$$\mathcal{A}^{\mathcal{S}}(f,\gamma) = pL - \bar{f} \qquad \forall (f,\gamma) \in \Gamma \subset \mathcal{S}; (\bar{f},1) \in \Gamma \tag{20}$$

$$\Delta f = \left[\max_{(f,\gamma)\in\Gamma} f\right] - \left[\min_{(f,\gamma)\in\Gamma} f\right] \tag{21}$$

$$\mathcal{D}(f,\gamma,f',\gamma') = |f - f'|/\Delta f + |\gamma - \gamma'| \qquad (f,\gamma) \in \mathcal{S}\setminus\Gamma, (f',\gamma') \in \Gamma \tag{22}$$

$$\mathcal{A}^{\mathcal{S}}(f,\gamma) = (pL - \bar{f})\left[1 - \min_{(f',\gamma')\in\Gamma} \mathcal{D}(f,\gamma,f',\gamma')\right] \qquad \forall (f,\gamma) \in \mathcal{S}\setminus\Gamma \tag{23}$$

- *(a) Parameter setting:* fix the initial population *popsize*, the number of generations *ng*, the rate *pr*1 of Rule1, the rate *pr*2 of Rule2, for each generation the number of clones *nc*, the mutation rate *mr*, the number of mutations *nm*, the number of exchangeable antibodies *nea*.
- *(b) Initial population generation:* create $pr1 \cdot popsize$ initial solutions by Rule1 and produce $pr2 \cdot popsize$ initial solutions by Rule2.

Step 2 - Objective function assessment:

- *(a)* calculate the objective function f in (18) and the feasibility degree γ in (19) for each antibody.
- *(b)* determine the Pareto set for (f,γ) pairs.
- *(c)* calculate the affinity for each antibody as reported in (20) and (23).

Step 3 - Clonal selection and expansion:

- *(a)* Take *nc* antibodies, from the population, with the greatest affinity.
- *(b)* Produce *nc* copies of the antibodies considered in Step 3a exploiting a binary tournament rule (randomly select two antibodies from *nc* antibodies and determine the best antibody).

Step 4 - Generating the next population:

- *(a)* Randomly choose *nm* antibodies from *nc* clones and use the mutations to create *nm* extra antibodies. Apply each mutation operator with the probability *mr*.
- *(b)* Include the *nm* extra antibodies to the current generation.
- *(c)* Exchange *nea* worst antibodies with new ones produced like those in Step 1b.
- *(d)* Copy the Pareto optimal antibodies to the next generation.

Step 5 - Conclusion test:

- *(a)* If the stopping criterion is met, return the Pareto optimal antibodies; otherwise, go to Step 2

At Step 1a, Rule1 is the full random rule: we randomly select one by one the nodes in the set $V \setminus \{n+1, n+p+1\}$. While in Rule2 we choose nodes using a probability that is inversely proportional to the distance between the current node and candidate node. Mutation operator randomly selects two solution indexes (Section 3.1) and swaps content. If the mutation makes unfeasible the depot node sequence $n+2,\ldots,n+p$, mutation is canceled.

Considering the ordinary AIA approach, the innovation in the method presented in this paper differs in the steps *2b*, *2c* and *4d*. Indeed, step *2b* and *2c* are used to determine the new antibody affinity. Whereas, step *4d* preserves the entire Pareto set in the next generation.

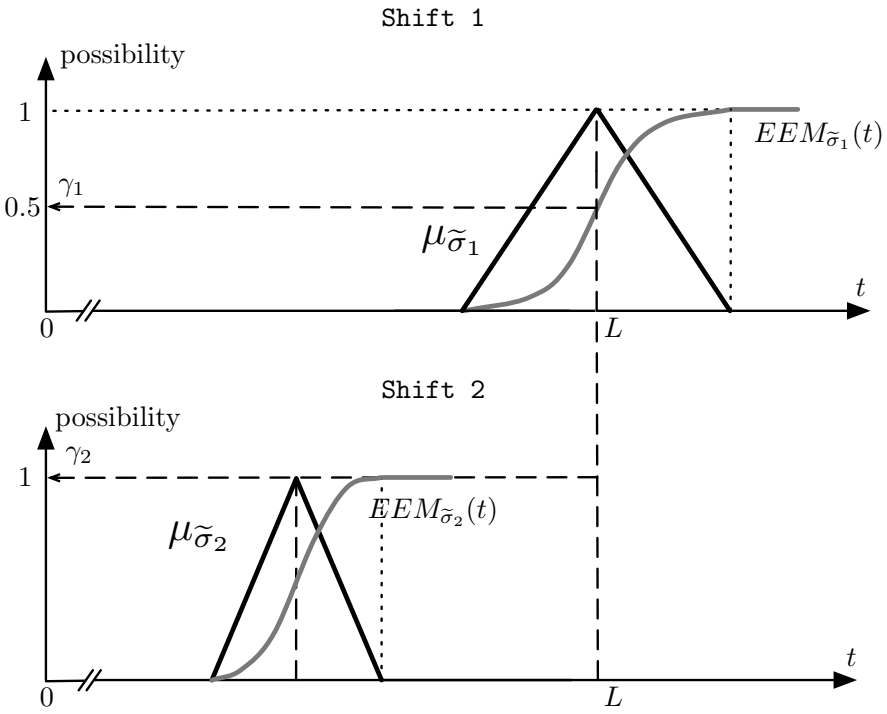


Figure 3: Fuzzy shift durations for solution Ψ

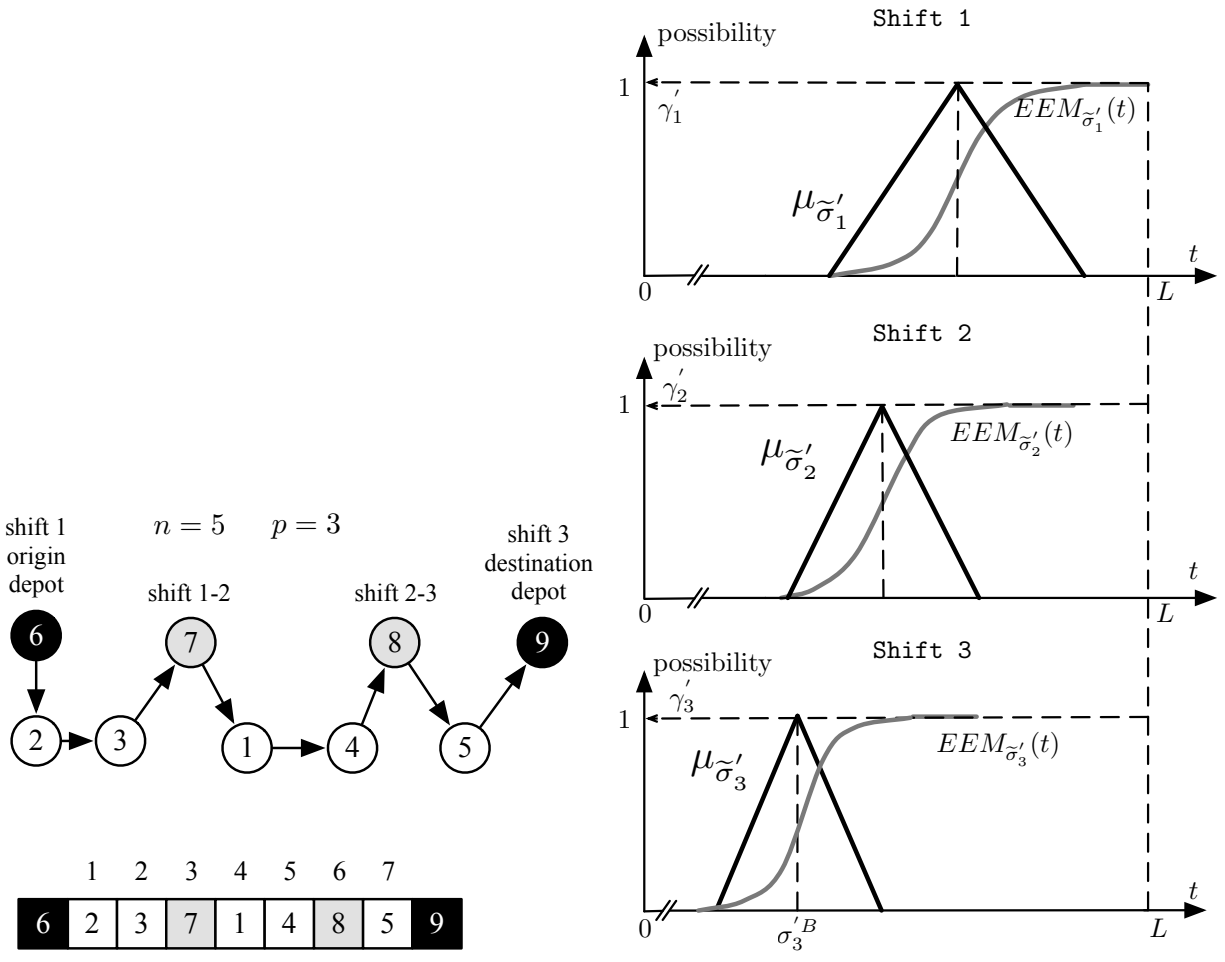


Figure 4: Solution Ψ' graph path, encoding and fuzzy shift durations

4 Numerical results

We assessed the performance of the approach described in Section 3.3 by generating different experiments. First, considering a standard non-fuzzy problem, we confronted the performance of the MIP reported in (1)-(13) in Section 2, with the AIA, that is Section 3.3 without steps *2b*, *2c*, and *4d*. Second, introducing fuzzy uncertainty, we adopted the AIA reported in Section 3.3 considering innovative steps *2b*, *2c*, and *4d*.

We developed an application on an x86 family 2.5 GHz Intel Core i7 processor having 4GB RAM and SDD, exploiting "C#.Net4" language. In the following, the AIA parameters and their criteria are defined. Population size $pop = 200$, number of generations $ng = 10000$, rate for Rule1 and Rule2 $pr1 = pr2 = 0.5$, number of clones $nc = 20$, mutation rate $mr = 0.75$, mutation number per generation $nm = 40$, exchangeable antibodies number $nea = 20$.

Three industrial test cases A, B, and C are adopted. In Table 3 the number of demand nodes n and the number of available shifts p is reported. Maximum shift duration is set to $L = 480$ min. Since industrial records are protected by a non-disclosure agreement, we report only summary data. The considered geographical is the Salento region in southeast Italy. Traveling and processing times are subject to uncertainty and are set considering user-experience values. Job processing times are equally distributed in three groups: small (around 15 min), medium (around 30 min) and large (around 40 min).

4.1 Non-fuzzy version results

Focusing on the non-fuzzy version of case studies A, B, and C, we compared theperformance of the MIP model in (1)-(13) with AIA. Crisp values t^B are considered in place of the fuzzy version $\tilde{t}$. The non-fuzzy MIP model reported in (1)-(13) was solved with software package IBM CPLEX v.12.5 setting a time limit equal to the running time of the AIA application. In Table 4 we described the achieved results. In particular, for each case study and for each approach, we indicated the time (sec) when a new better solution is detected together with the corresponding solution optimal gap (%). In the first case study A, AIA calculates formerly the MIP optimal solution. For case study B, the optimal solution gap of AIA from MIP is 1% but AIA is tenfold faster than MIP. Finally, in the case study C, the optimal gap is 4%.

4.2 Fuzzy version results

Considering the fuzzy version of case studies A, B, and C, we adopted the fuzzy AIA to obtain the solution Pareto set. As reported in Section 3, multiple optimal solutions exist because of the 2-factor comparison. Company manager receives the optimal Pareto set and assesses the best solution. In Section 4.2.1, an example is provided. Assuming different AIA parameter values leads to different Pareto sets as shown in Section 4.2.2.

Our main objective consists in determining the AIA parameters to find as many Pareto set solutions. Consequently, we designed an experimental campaign to determine the best combination of AIA parameter values, see Section 4.2.3. Each experiment ϵ produces a solution Pareto set Γ_ϵ. Considering all the campaign experiments $\mathcal{E} = \{1, \ldots, exp\}$ leads merging the corresponding Pareto sets $\bigcup_{\epsilon \in \mathcal{E}} \Gamma_\epsilon$. Union of Pareto sets is not a Pareto set, because a solution of a Pareto set can be dominated by a solution of another Pareto set experiment. We define $\overline{\Gamma}^{\mathcal{E}}$ as Hyper Pareto set for experiments $1, \ldots, exp$ containing only the dominant solutions obtained from the experimental campaign. For each experiment ϵ, we denote the experiment impact ρ_ϵ as the number of solutions belonging both to the experiment Pareto set Γ_ϵ and the global Hyper Pareto set $\overline{\Gamma}^{\mathcal{E}}$.

Table 3: Case study definition

Case study	n	p
A	21	5
B	33	8
C	45	10

4.2.1 Single experiment results

Considering the fuzzy version of case study A, our software produces the Pareto set reported in Fig. 5. The collected solutions $s1, s2, \ldots s10$, belonging to the optimal Pareto set, are reported in Table 5. For such solutions, jobs are always completed in 3 shifts.

Feasibility degree for solution $s1$ is 1, because $\sigma_i^C < 480$, $\forall i = 1, 2, 3$. Solution $s1$ makespan is $2 \cdot 480 + 348 = 1308$ min. Since solution $s1$ is very conservative, the manager is unlikely to accept such a high safety margin.

Whereas, solution $s2$ has $f = 1263$ min and $\gamma = 0.968$. Indeed, in Table 5 we have $\sigma_1^C = 486 > 480$ and $\sigma_2^C = 499 > 480$; we note that 96.8% of the area of fuzzy number $\widetilde{\sigma}_2 = (349; 424; 499)$ stays on the left side of 480 (see Fig. 1). The manager prefers solution $s2$ to $s1$ as assuming a low risk (3.2%) he/she decreases makespan by 45 minutes. Solutions $s3$ and $s4$ have similar feasibility degrees but $s4$ makespan is lower than $s3$. Whereas, solutions s5, s6, and s7 have similar makespan but $s5$ feasibility degree is better than others. Feasibility degrees for solutions s8, s9, and s10 are lower than 0.5: see $\sigma_2^B > 480$ min in last three rows in Table 5.

Referring to Fig. 5, manager declares that solutions $s2$, $s4$, and $s5$ are the most valuable and selects the best one based on his/her experience.

In Fig. 6, shift tours for solution $s1$ are graphically reported using 'Google Maps' website (maps.google.com). For privacy agreement, only the position of Lecce, Salento main city, is indicated. Depot position is red pinned. Job locations are indicated with circles. Basically, shift 1 tour accomplishes jobs in the north side. Shift 2 serves central zone and shift 3 satisfies south side. Each shift serves 7 jobs.

Table 4: Comparison AIA vs. MIP for non-fuzzy version of case studies A, B, C

Case study A				Case study B				Case study C			
AIA		MIP		AIA		MIP		AIA		MIP	
t [sec]	gap [%]	t [sec]	gap [%]	t [sec]	gap [%]	t [sec]	gap [%]	t [sec]	gap [%]	t [sec]	gap [%]
1	130	1	230	1	123	1	450	1	187	1	345
20	95	31	99	31	115	30	129	21	125	36	234
45	43	76	87	51	33	45	97	48	49	54	221
105	22	273	54	120	12	150	65	532	4	169	198
123	0	778	23	246	1	264	51			254	187
		1002	13			778	42			675	85
		1324	6			1042	21			1201	31
		1457	0			1256	14			1312	20
						2001	9			1416	12
						2398	2			2395	7
						2405	0			5998	2
										7194	0

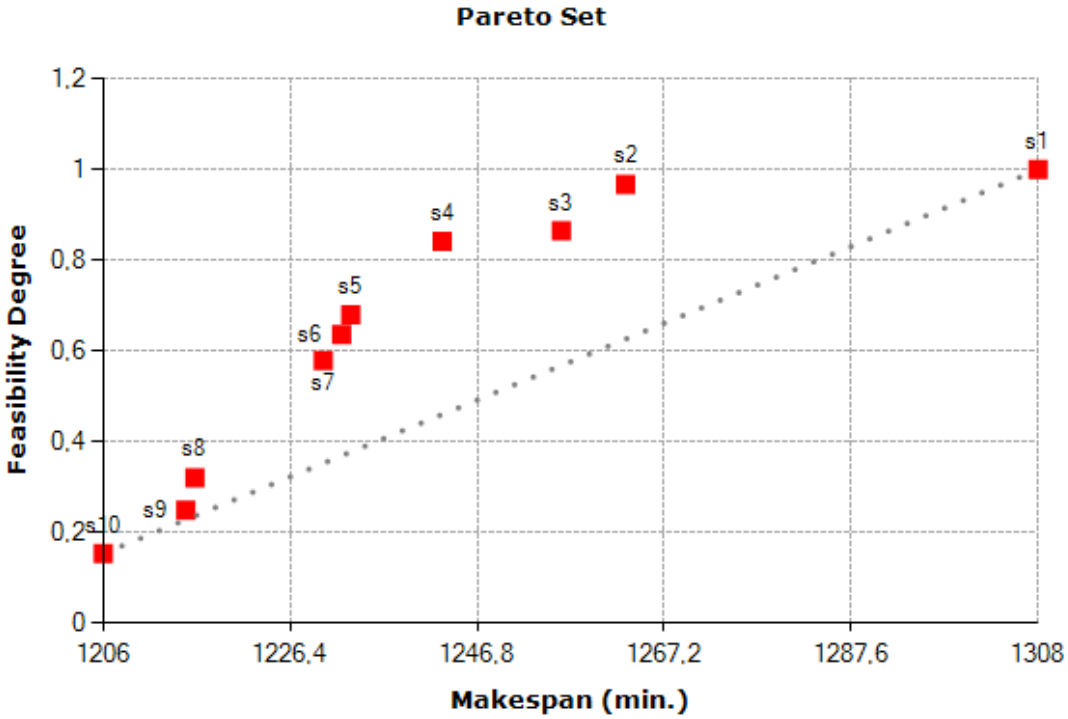
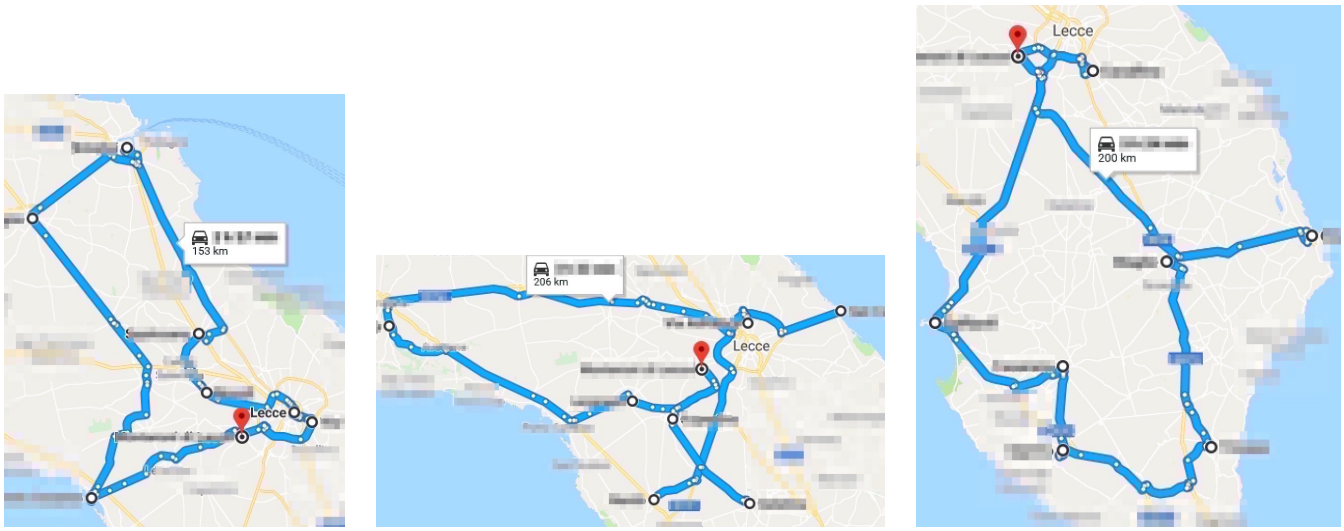


Figure 5: Final solution Pareto set for fuzzy case study A with basic parameters

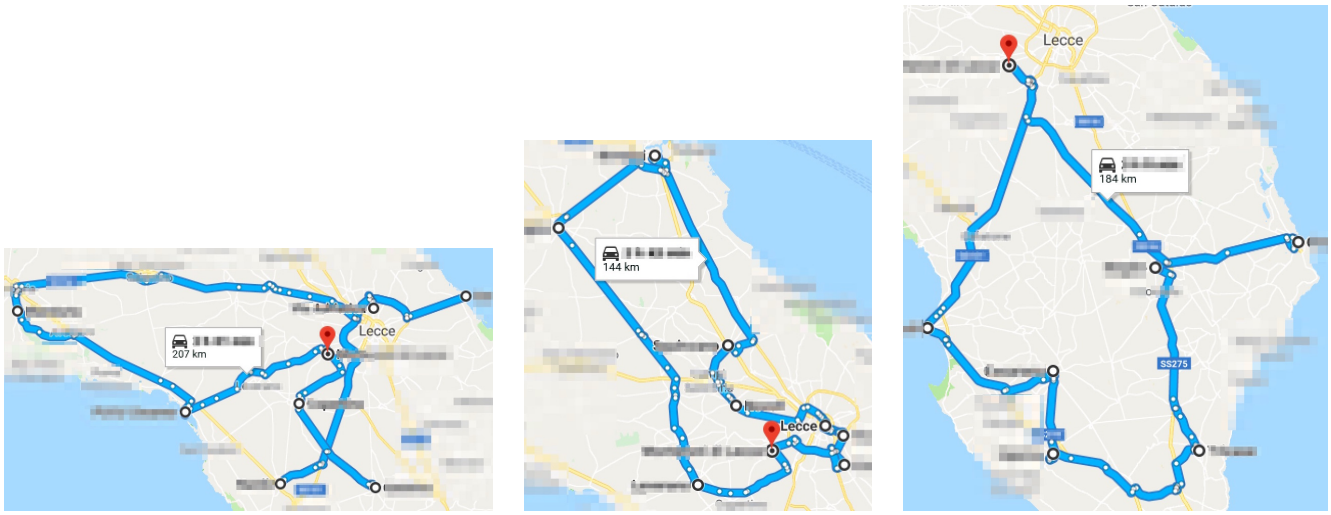
Table 5: Details of final solutions for fuzzy case study A with basic parameters

	Actual Fuzzy Shift Duration $\widetilde{\sigma_i}$		
Solution	**Shift 1**	**Shift 2**	**Shift 3**
$s1$	(357;417;477)	(345;410;475)	(298;348;398)
$s2$	(366;426;486)	(349;424;499)	(263;303;343)
$s3$	(372;432;492)	(369;444;519)	(256;296;336)
$s4$	(390;450;510)	(365;445;525)	(248;283;318)
$s5$	(402;467;532)	(369;444;519)	(238;273;308)
$s6$	(402;467;532)	(394;469;544)	(237;272;307)
$s7$	(400;460;520)	(388;473;558)	(240;270;300)
$s8$	(400;460;520)	(412;497;582)	(226;256;286)
$s9$	(418;473;528)	(420;505;590)	(220;255;290)
$s10$	(371;426;481)	(433;518;603)	(211;246;281)



Shift 1 Shift 2 Shift 3

Figure 6: Shift tours for the solution $s1$



Shift 1 Shift 2 Shift 3

Figure 7: Shift tours for the solution $s2$

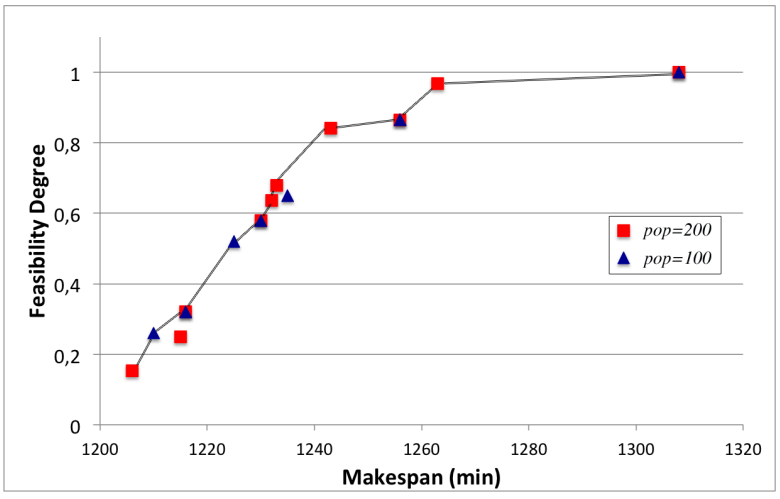


Figure 8: Final solution Pareto set for fuzzy case study A with two experiments

Table 6: Experiment impact results

(a) Case study A

exp	*ng*	*pop*	*pr*1	ρ
1	5000	100	0.25	5
2	5000	100	0.50	5
3	5000	100	0.75	5
4	5000	200	0.25	7
5	5000	200	0.50	8
6	5000	200	0.75	6
7	5000	300	0.25	7
8	5000	300	0.50	8
9	5000	300	0.75	6
10	10000	100	0.25	5
11	10000	100	0.50	6
12	10000	100	0.75	5
13	10000	200	0.25	8
14	10000	200	0.50	9
15	10000	200	0.75	8
16	10000	300	0.25	8
17	10000	300	0.50	9
18	10000	300	0.75	9
19	15000	100	0.25	5
20	15000	100	0.50	6
21	15000	100	0.75	5
22	15000	200	0.25	8
23	15000	200	0.50	9
24	15000	200	0.75	9
25	15000	300	0.25	8
26	15000	300	0.50	9
27	15000	300	0.75	9

(b) Case study B

exp	*ng*	*pop*	*pr*1	ρ
1	5000	100	0.25	7
2	5000	100	0.50	7
3	5000	100	0.75	7
4	5000	200	0.25	8
5	5000	200	0.50	11
6	5000	200	0.75	8
7	5000	300	0.25	9
8	5000	300	0.50	11
9	5000	300	0.75	8
10	10000	100	0.25	7
11	10000	100	0.50	7
12	10000	100	0.75	7
13	10000	200	0.25	10
14	10000	200	0.50	12
15	10000	200	0.75	11
16	10000	300	0.25	11
17	10000	300	0.50	12
18	10000	300	0.75	12
19	15000	100	0.25	7
20	15000	100	0.50	8
21	15000	100	0.75	7
22	15000	200	0.25	11
23	15000	200	0.50	13
24	15000	200	0.75	12
25	15000	300	0.25	11
26	15000	300	0.50	13
27	15000	300	0.75	14

(c) Case study C

exp	*ng*	*pop*	*pr*1	ρ
1	5000	100	0.25	9
2	5000	100	0.50	9
3	5000	100	0.75	9
4	5000	200	0.25	9
5	5000	200	0.50	14
6	5000	200	0.75	11
7	5000	300	0.25	12
8	5000	300	0.50	15
9	5000	300	0.75	11
10	10000	100	0.25	10
11	10000	100	0.50	9
12	10000	100	0.75	10
13	10000	200	0.25	13
14	10000	200	0.50	16
15	10000	200	0.75	16
16	10000	300	0.25	15
17	10000	300	0.50	16
18	10000	300	0.75	16
19	15000	100	0.25	10
20	15000	100	0.50	11
21	15000	100	0.75	10
22	15000	200	0.25	15
23	15000	200	0.50	18
24	15000	200	0.75	16
25	15000	300	0.25	15
26	15000	300	0.50	18
27	15000	300	0.75	17

In Fig. 7, tours for solution $s2$ are reported. The comparison of solution $s2$ with $s1$ is described in the following (see Fig. 6 and 7). Shift 1 tour lies in the central area whereas shift 2 concerns the north zone. In $s2$, to reduce makespan, one job is removed from shift 3 and added to shift 2. A job swap is performed between shift 1 and 2 in order to limit the shift 2 duration.

4.2.2 Two-experiment results

We compare the experiment performed in previous section 4.2.1 ($\epsilon = 1$), with a new experiment ($\epsilon = 2$) in which *pop* parameter is decreased from 200 to 100. Consequently, considering the experimental campaign set $\mathcal{E} = \{1, 2\}$, we report the corresponding solution Pareto sets Γ_1 ($pop = 200$) and Γ_2 ($pop = 100$) in Fig. 8. Set Γ_1, in red, is reported in Fig. 5 too. Set Γ_2 contains 7 solutions, in particular: 4 solutions belong to Γ_1 too, 2 solutions are dominant solutions not included in Γ_1 and one solution is dominated by another in Γ_1. Instead, one solution in Γ_1 is dominated by a solution in Γ_2. Since, the new experiment ($\epsilon = 2$) produces two additional solutions in Pareto set and removes one, we have $|\overline{\Gamma}^{\mathcal{E}}| = 11$. In Fig. 8, the line represents Pareto front. Calculating $\rho_1 = 9$ and $\rho_2 = 6$, we can assess that first experiment has a greater impact than second.

4.2.3 Experimental campaign results

Considering the case study described in Section 4, we designed an experimental campaign. We examined 3 parameters *ng*, *pop*, and *pr*1 on 3 different levels: $ng \in \{5000, 10000, 15000\}$, $pop \in \{100, 200, 300\}$, and $pr1 \in \{0.25, 0.50, 0.75\}$. Consequently, for each case study, we have 27 experiments, $\mathcal{E} = \{1, \ldots, 27\}$. For each experiment $\epsilon \in \mathcal{E}$, we calculate the corresponding impact ρ_ϵ. In Table 6a, experiment impact values are reported for case study A. As parameter *pop* increases, we obtain a larger number of Pareto set solutions. Increasing the parameter *ng* from 5000 to 10000 succeeds in increasing ρ, the same cannot be said when passing from 10000 to 15000. Parameter *pr*1 has a positive influence when $pr1 = 0.50$.

In Table 6b and Table 6c, experiment impact values are reported for case study B and C. The results confirm the findings of case study A. In conclusion, our solution Pareto set approach significantly depends on population parameter *pop*. Indeed, in each generation, we preserve the 'entire' solution Pareto set, so it is important having a larger *pop* value than the non-fuzzy approach. Parameter $pr1 = pr2 = 0.50$ represents a good compromise between generating Rule1 full random solutions and Rule2 solutions (Section 3.3).

5 Conclusion

This study presents some real insights into the single-vehicle routing problem with multi-shift and fuzzy uncertainty. Our objective consists of minimizing both the system makespan and shift overtime occurrence. In particular, we investigated the effect of uncertainty in driving and job processing time. We provide optimal solutions for the decision-maker considering a 2-factor comparison: the objective function value (makespan) and the degree with which overtime is avoided. Our approach is currently used by the case study company.

References

[1] S. Zangeneh-Khamooshi, Z. B. Zabinsky, and J. A. Heim, "A multi-shift vehicle routing problem with windows and cycle times," *Optimization Letters*, vol. 7, no. 6, pp. 1215–1225, aug 2013.

[2] Y. Ren, M. Dessouky, and F. Ordóñez, "The multi-shift vehicle routing problem with overtime," *Computers & Operations Research*, vol. 37, no. 11, pp. 1987–1998, 2010.

[3] G. Onder, I. Kara, and T. Derya, "New integer programming formulation for multiple traveling repairmen problem," *Transportation Research Procedia*, vol. 22, pp. 355–361, 2017. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S2352146517301771

[4] S. Nucamendi-Guillén, I. Martínez-Salazar, F. Angel-Bello, and J. M. Moreno-Vega, "A mixed integer formulation and an efficient metaheuristic procedure for the k-Travelling Repairmen Problem," *Journal of the Operational Research Society*, vol. 67, no. 8, pp. 1121–1134, aug 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1057/jors.2015.113

[5] K. Sparks, C. L. Cooper, Y. Fried, and A. Shirom, "The Effects of Working Hours on Health: A Meta-Analytic Review," in *From Stress to Wellbeing Volume 1*. London: Palgrave Macmillan UK, 2013, pp. 292–314.

[6] C. C. Caruso, "Negative Impacts of Shiftwork and Long Work Hours," *Rehabilitation Nursing*, vol. 39, no. 1, pp. 16–25, jan 2014.

[7] G. Costa, "Shift Work and Health: Current Problems and Preventive Actions," *Safety and Health at Work*, vol. 1, no. 2, pp. 112–123, 2010.

[8] T. M. Cook and R. A. Russell, "A Simulation and Statistical Analysis of Stochastic Vehicle Routing with Timing Constraints," *Decision Sciences*, vol. 9, no. 4, pp. 673–687, oct 1978.

[9] M. Gendreau, G. Laporte, and R. Séguin, "Stochastic vehicle routing," *European Journal of Operational Research*, vol. 88, no. 1, pp. 3–12, jan 1996.

[10] G. Yaohuang, X. Binglei, and G. Qiang, "Overview of Stochastic Vehicle Routing Problems," *Journal of Southwest Jiaotong University*, vol. 10, no. 2, pp. 113–121, 2002.

[11] C. Lee, K. Lee, and S. Park, "Robust vehicle routing problem with deadlines and travel time/demand uncertainty," *Journal of the Operational Research Society*, vol. 63, no. 9, pp. 1294–1306, sep 2012.

[12] J. Xu, F. Yan, and S. Li, "Vehicle routing optimization with soft time windows in a fuzzy random environment," *Transportation Research Part E: Logistics and Transportation Review*, vol. 47, no. 6, pp. 1075–1091, 2011.

[13] H. Ewbank, P. Wanke, and A. Hadi-Vencheh, "An unsupervised fuzzy clustering approach to the capacitated vehicle routing problem," *Neural Computing and Applications*, vol. 27, no. 4, pp. 857–867, may 2016.

[14] Z. Zhu, J. Xiao, S. He, Z. Ji, and Y. Sun, "A multi-objective memetic algorithm based on locality-sensitive hashing for one-to-many-to-one dynamic pickup-and-delivery problem," *Information Sciences*, vol. 329, no. C, pp. 73–89, feb 2016.

[15] A. G. Novaes, E. T. Bez, P. J. Burin, and D. P. Aragão, "Dynamic milk-run OEM operations in over-congested traffic conditions," *Computers & Industrial Engineering*, vol. 88, no. C, pp. 326–340, oct 2015.

[16] H. Zhao, W. A. Xu, and R. Jiang, "The Memetic algorithm for the optimization of urban transit network," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3760–3773, may 2015.

[17] D. Muñoz-Carpintero, D. Sáez, C. E. Cortés, and A. Núñez, "A Methodology Based on Evolutionary Algorithms to Solve a Dynamic Pickup and Delivery Problem Under a Hybrid Predictive Control Approach," *Transportation Science*, vol. 49, no. 2, pp. 239–253, may 2015.

[18] S. F. Ghannadpour, S. Noori, R. Tavakkoli-Moghaddam, and K. Ghoseiri, "A multi-objective dynamic vehicle routing problem with fuzzy time windows: Model, solution and application," *Applied Soft Computing*, vol. 14, pp. 504–527, jan 2014.

[19] D. Sáez, C. E. Cortés, and A. Núñez, "Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering," *Computers & Operations Research*, vol. 35, no. 11, pp. 3412–3438, nov 2008.

[20] L. de C.T. Gomes and F. J. Von Zuben, "Multiple criteria optimization based on unsupervised learning and fuzzy inference applied to the vehicle routing problem," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 13, no. 2-4, pp. 143–154, 2002.

[21] Y. He and J. Xu, "A class of random fuzzy programming model and its application to vehicle routing problem," *World Journal of Modelling and simulation*, 2005.

[22] D. Teodorović and G. Pavković, "The fuzzy set theory approach to the vehicle routing problem when demand at nodes is uncertain," *Fuzzy Sets and Systems*, vol. 82, no. 3, pp. 307–317, sep 1996.

[23] K. Tan and K. Tang, "Vehicle dispatching system based on Taguchi-tuned fuzzy rules," *European Journal of Operational Research*, vol. 128, no. 3, pp. 545–557, 2001.

[24] M. Jiménez, M. Arenas, A. Bilbao, and M. V. Rodrguez, "Linear programming with fuzzy parameters: An interactive method resolution," *European Journal of Operational Research*, vol. 177, no. 3, pp. 1599–1609, mar 2007.

[25] H.-J. Zimmermann, M. A. E. Kassem, N. M. El-Badry, and A. Bilbao, "Fuzzy mathematical programming," *Computers & Operations Research*, vol. 10, no. 4, pp. 291–298, jan 1983.

[26] M. Jiménez, "Ranking Fuzzy Numbers Through The Comparison Of Its Expected Intervals," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 04, no. 04, pp. 379–388, aug 1996.

[27] H. J. J. Kals, International Design Seminar 1999 Enschede, and International Institution for Production Engineering Research, *Integration of process knowledge into design support systems : proceedings of the 1999 CIRP International Design Seminar, University of Twente, Enschede, The Netherlands, 24 - 26 March 1999*. Kluwer, 1999.

[28] Baoding Liu and Yian-Kui Liu, "Expected value of fuzzy variable and fuzzy expected value models," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 4, pp. 445–450, aug 2002.

[29] Quan Nguyen and Tu Van Le, "Fuzzy discrete-event simulation with FTipLog," in *1996 Australian New Zealand Conference on Intelligent Information Systems. Proceedings. ANZIIS 96*. IEEE, pp. 195–198.

[30] F. Nucci, "The multi-shift single-vehicle routing problem under fuzzy uncertainty," in *2017 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. IEEE, sep 2017, pp. 156–161. [Online]. Available: http://ieeexplore.ieee.org/document/8120987/

[31] M. Mobini, Z. Mobini, and M. Rabbani, "An Artificial Immune Algorithm for the project scheduling problem under resource constraints," *Applied Soft Computing*, vol. 11, no. 2, pp. 1975–1982, mar 2011.

[32] A. Bagheri, M. Zandieh, I. Mahdavi, and M. Yazdani, "An artificial immune algorithm for the flexible job-shop scheduling problem," *Future Generation Computer Systems*, vol. 26, no. 4, pp. 533–541, apr 2010.

[33] J. Al-Enezi, M. Abbod, and S. Alsharhan, "Artificial Immune Systems - models, algorithms and applications," *International Journal of Research and Reviews in Applied Sciences*, vol. 3, no. 2, pp. 118–131, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 36-44 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

On The Development of a Reliable Gate Stack for Future Technology Nodes Based on III-V Materials

Abhitosh Vais[1,*], Sonja Sioncke[1], Jacopo Franco[1], Vamsi Putcha[1,2], Laura Nyns[1], Arturo Sibaja-Hernandez[1], Rita Rooyackers[1], Sergio Calderon Ardila[2], Valentina Spampinato[1], Alexis Franquet[1], Jan Maes[3], Qi Xie[3], Michael Givens[4], Fu Tang[4], Xiang Jiang[4], Marc Heyns[1,2], Dimitri Linten[1], Jerome Mitard[1], Aaron Thean[5], Dan Mocuta[1] and Nadine Collaert[1]

[1]*IMEC, Kapeldreef 75, 3001 Leuven, Belgium*

[2]*ESAT department, KU Leuven, 3000 Leuven, Belgium*

[3]*ASM Belgium, Kapeldreef 75, 3001 Leuven, Belgium*

[4]*ASM America, Phoenix, AZ 85034, USA*

[5]*Department of Electrical and Computer Engineering, NUS, 117546 Singapore*

ARTICLE INFO

Article history:
Received: 29 May, 2018
Accepted: 02 July, 2018
Online: 15 September, 2018

Keywords:
InGaAs
III-V passivation
III-V reliability
III-V MOSFET
High-κ oxides
Oxide defects

ABSTRACT

In this work, we discuss how the insertion of a $LaSiO_x$ layer in between an in-house IL passivation layer and the high-k has moved the III-V gate stack into the target window for future technology nodes. The insertion of this $LaSiO_x$ layer in the gate stack has reduced the D_{it} and N_{bt} below the target level of $5x10^{11}$ /eV.cm^2 and $3x10^{10}$ /cm^2 (target at 10 years operation: ΔV_{fb}<30mV at 125°C) respectively. From physical analysis, it was found that $LaSiO_x$ can stabilize the interaction of the IL layer with the InGaAs substrate. An implant free $In_{0.53}Ga_{0.47}As$ n-MOSFET was fabricated with this gate stack and for the first time, a III-V gate stack meets the reliability target for advanced technology nodes with a max operating V_{ov} of 0.6 V. In addition, an excellent electron mobility ($\mu_{eff,peak}$=3531 cm^2/V-s), low SS_{lin}=71 mV/dec and an EOT of 1.15 nm were obtained. We also report the scaling potential of this stack to 1 nm EOT without any loss in performance, reliability and further reduction of the sub-threshold swing (SS_{lin}=68 mV/dec).

1. Introduction

III-V compound semiconductors have already been widely investigated as an alternative material to Silicon for future Metal - Oxide-Semiconductor Field Effect Transistors (MOS-FETs). MOS devices based on III-V compound semiconductor materials like InGaAs, with excellent performance, have been demonstrated on both 2-inch InP substrates and on a 300mm Si platform [1-5]. But, in order to become a commercially viable alternative to existing Si technologies, two major issues are yet to be solved: 1) the defective interface between III-V materials and the high-k oxide layers, and 2) the high density of defects in the most commonly investigated high-k oxides, Al_2O_3 and HfO_2.

Both of these issues have been shown to affect almost every aspect of device performance, such as the capacitance [6-11], reliability of III-V based devices [11-14], on-state electrical parameters like mobility [15], and trans-conductance [16, 17]. As a result, it becomes extremely important to reduce their impact in order to make commercial III-V MOSFET technology a reality. In this regard, the use of a new IL layer in combination with an Al_2O_3/HfO_2 or just HfO_2 stack was reported, showing improved performance and a significant reduction of the oxide traps. However, with the requirement of a thin Al_2O_3 layer for gate stack thermal stability and due to the wide defect distribution in Al_2O_3, the total density of charging oxide traps was still higher than the target (10 years reliability target: flat-band voltage shift, $\Delta V_{fb} < 30$ mV at 125°C. Assuming a typical time power law exponent of 0.13, the failure criterion projects effective density of charged defects, ΔN_{eff} to < $3x10^{10}$/cm^2 at operating field, for an Effective Oxide Thickness (EOT) of 1 nm [18, 19]).

*Abhitosh Vais, Email: Abhitosh.Vais@imec.be

This paper is an extension of the work presented in *IEEE Symposium on VLSI Technology 2017*. In this paper, we demonstrate and comprehensively discuss on the process of replacing Al_2O_3 in the IL/Al_2O_3/HfO_2 gate stack with $LaSiO_x$. Furthermore, we elaborate on the improvements, thereof, in the interface properties and reliability for operation at V_{DD}=0.8 V, which exceeds the operating target for future technology nodes without any loss in the electrical performance of the MOSFET device reported earlier.

2. Experimental details

Capacitors were made starting from an MBE grown 300 nm n-type $In_{0.53}Ga_{0.47}As$ on 2-inch n-type InP substrates. The targeted Si doping in the $In_{0.53}Ga_{0.47}As$ grown layer was ~$1x10^{17}$ at/cm^3. Prior to the gate stack deposition, the substrates were cleaned in a 2M HCl solution for 5 min at room temperature and subsequently rinsed in de-ionized water. H_2S pretreatment and deposition of the IL (Inter Layer), $LaSiO_x$, Al_2O_3 and HfO_2 were done in an ASM Pulsar® 3000 ALD reactor. The H_2S pretreatment is done in-situ prior to the high-κ deposition. The IL used here is an ALD based layer with a κ-value ~6. The H_2S treatment and IL film deposition are both done at 250°C. Al_2O_3 and HfO_2 were deposited by ALD TMA/H_2O and $HfCl_4$/H_2O at 300°C. $LaSiO_x$ is deposited by an ALD process at 250°C using the following precursors: LoLaPrime® (AIR Liquide), $SiCl_4$, H_2O. A 42% La concentration in the $LaSiO_x$ is obtained by using a La:Si pulse ratio of 1:5. A full ALD cycle consists of : [[$SiCl_4$/purge/H2O/purge] x 5 + [LolaPrime/purge/H_2O/purge] x 1] x n. Under these conditions, a GPC of 0.1754 nm is achieved.

Dot capacitors were fabricated by sputtering 80 nm TiN using physical vapor deposition (PVD). A dry-etch process was used to etch the metal gate. An anneal in forming gas (10% H_2/N_2) was done prior to metal deposition (PDA) or post metal deposition (PMA) at 400°C for 5 min.

Implant-free quantum-well devices were also fabricated to study the impact of these passivation schemes on the device performance and reliability. A schematic diagram of the quantum-well MOSFET is shown in Fig.1. The stack for MOSFETs consists of a 15 nm un-intentionally doped $In_{0.53}Ga_{0.47}As$ channel layer, a 3 nm InP etch stop layer and a 50 nm n+ $In_{0.53}Ga_{0.47}As$ (Si-doped, $1x10^{19}$ cm^{-3}) layer, all grown on a 2-inch semi-insulating InP substrate. Fig. 2 describes the process flow used in this work. The detailed device fabrication can be found in [20]. Prior to gate stack deposition, a digital etch (cycles of HCl and H_2O_2) was used to clean and smoothen the InGaAs channel surface. The different gate stacks that were studied are depicted in Table I.

I-V and C-V measurements were used to study the impact of the gate stack on the performance of the devices. I-V measurements were carried out using HP4156 parameter analyzer and C-V measurements were performed using a HP4284A-LCR meter. The interface defect density, D_{it}, is extracted using the conductance method from the parallel equivalent conductance G_p [21] at room temperature. The amount of oxide traps (ΔN_{eff}) are measured by C-V hysteresis measurements (forward and reverse sweep of the gate voltage) at different maximum stress voltages and the corresponding shift in flat band voltage, ΔV_{fb}, measured between the forward and backward trace is converted into a sheet charge (ΔN_{eff} (E_{ox})). The flat-band voltage is the gate voltage corresponding to an inflection point in the double derivative of the normalized capacitance-voltage data at a frequency of 100 kHz [21,22]. The ΔN_{eff} at target field as well as the voltage acceleration

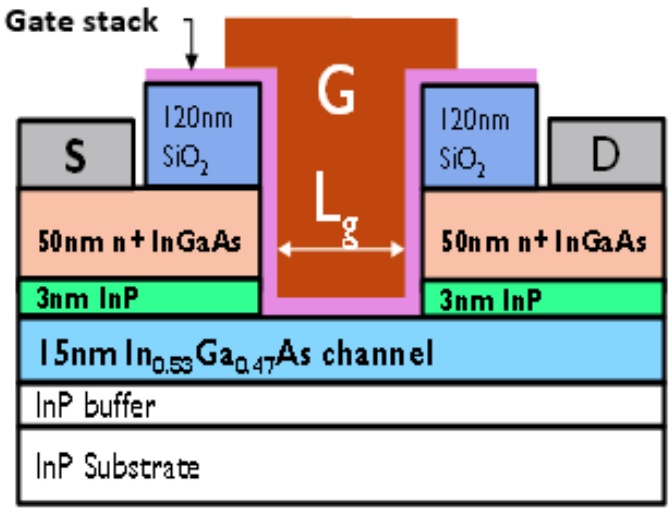


Figure. 1. Schematic representation of the MOSFET device (Dimensions are not to scale).

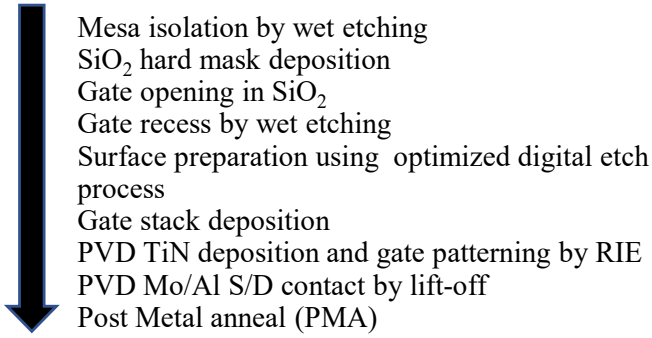


Figure 2. Process flow used in this work for MOSFET fabrication.

Table 1: Split table of III-V gate stacks

Sample ID	Dielectric
Ref	3 nm HfO_2
A	1 nm IL/3 nm HfO_2
A'	0.5 nm IL/3 nm HfO_2
B	1 nm IL/1 nm Al_2O_3/3 nm HfO_2
B'	1 nm IL/0.2 nm Al_2O_3/3 nm HfO_2
C	1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2
C'	1 nm $LaSiO_x$/3 nm HfO_2

factor (γ, i.e. the slope of the log-log plot of ΔN_{eff} vs E_{ox}) is reported as a figure of merit to study the oxide trap behavior [23]. Although this method does not take into account the influence of stress temperature and stress time into account, it helps in identifying the trends between different gate stack/pretreatments. Bias temperature instability (BTI) characterization using extended Measure-Stress-Measure (eMSM) technique was used to evaluate the reliability of the MOSFET devices [23]. Finally, Time-Of-Flight Secondary Ion Mass Spectrometry (TOFSIMS) was used to study the chemistry of the stacks.

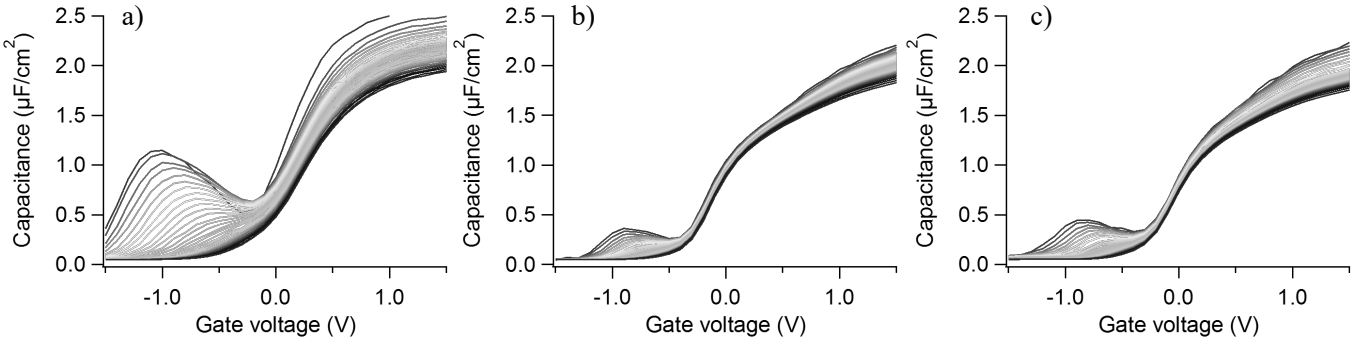


Figure 3. Measured RT CV's of gate stacks a) 'ref', b) A and c) B' described in Table I at 31 frequencies between 1kHz and 1MHz. The MOSCAP devices received a PDA at 400°C, 5' in Forming gas.

3. Results and discussion

3.1. The IL layer development and thermal stability

Fig. 3 shows the multi-frequency (1 kHz - 1 MHz) C-V for various samples in the split table of Table I, measured at room temperature. The reduction in the D_{it} bump and in frequency dispersion when inserting the IL (3-A,B') between the InGaAs channel and the HfO_2 layer clearly shows the improvement. Fig. 4 shows the result of the extraction of various electrical parameters from the measured C-V data. The extracted mid-gap D_{it} drops by more than a factor of 2 when inserting the IL accompanied with a negative flat band voltage, V_{fb}, shift (Fig. 4a,b). An H_2S treatment prior to the HfO_2 deposition has a similar effect: it reduces the D_{it} and shifts the V_{fb} by -0.15 V. Compared to an H_2S treatment, the insertion of 1 nm of the IL layer shifts the V_{fb} to even more negative values and a larger reduction in D_{it} is observed. In addition, the IL layer reduces the ΔN_{eff} by 1 order of magnitude and increases the γ-value to 2.5. Although 0.5 nm of the IL layer shows similar trends, an optimum is reached when using 1nm of the IL layer. This result suggests that 1 nm of IL is needed to form a closed layer passivating the interface. When adding a thin (0.2 nm) Al_2O_3 cap on top of the IL layer, the D_{it} bump and frequency dispersion increase slightly (Fig. 3-B'). One can also observe that the V_{fb} shifts to more positive values and this goes hand in hand with a slight increase in ΔN_{eff} and reduction of the γ value. From these observations, a hypothesis can be put forward. Fig. 5 depicts

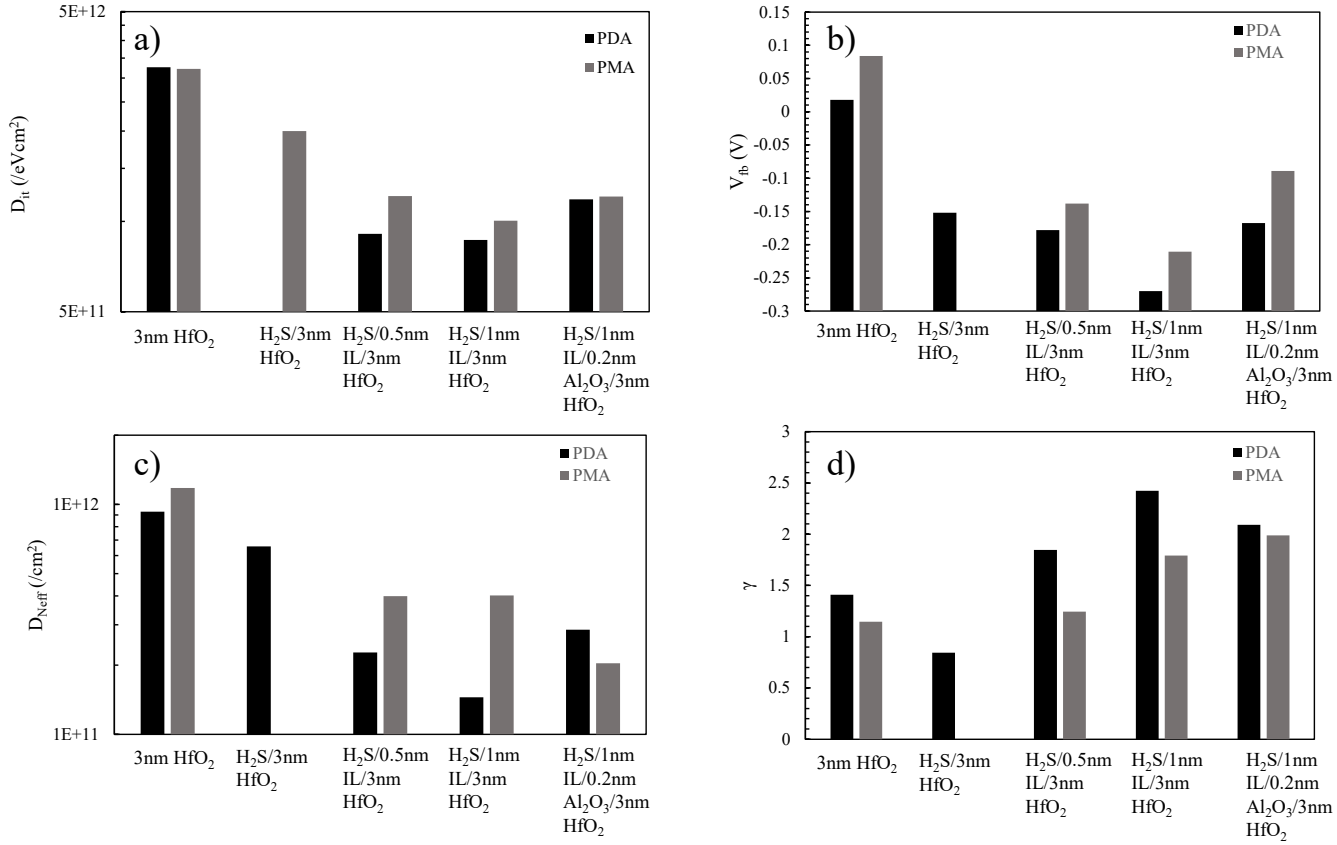


Figure 4. Electrical parameters extracted for gate stacks 'ref', A, A', B' described in Table I. a) Midgap D_{it} extracted from RT CV's. b) V_{fb} c) ΔN_{eff} extracted from RT CV hysteresis by converting the measured ΔV_{fb} into an effective charge sheet at the interface by: $\Delta N_{eff}=\Delta V_{fb}*C_{ox}/q$, and reported at charging equivalent field of 3.5 MV/cm [13]. V_{start} condition was taken at V_g=-1.5V. d) Field acceleration exponent, γ, i.e. the slope of the log-log plot ΔN_{eff} vs oxide field which is a measure of the misalignment between the channel carrier energy-i.e. InGaAs E_f- and defect levels in the high-*k*. Filled bars represent the data after PDA, shaded bars represent the data after PMA.

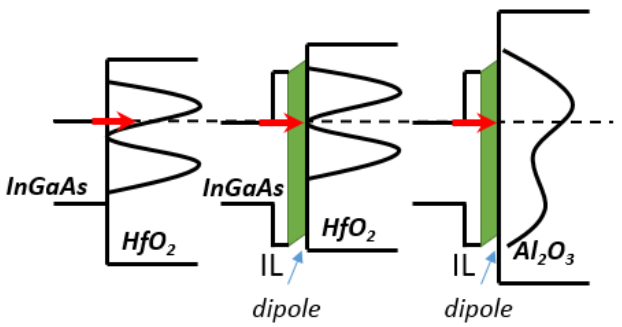


Figure 5. Defect band model to explain the observed changes in the γ value and total amount of charged oxide traps at a given field. The IL passivation layer is creating a dipole shifting the misalignment between InGaAs E_f and defect levels in the high-κ. In Al_2O_3 a partial overlap between deep and shallow traps is existing [9] and an almost uniform defect band is present. Therefore, the dipole created is not as effective as compared to the situation where deep and shallow traps are well separated as is proposed for the stack with HfO_2.

a cartoon of a possible explanation. Defect band modeling [24,25] has shown that although peak defect densities in Al_2O_3 are lower, a partial overlap of the defect bands induces an almost uniform distribution of defect levels around the InGaAs conduction band edge, E_c. On the contrary, HfO_2 on InGaAs shows a minimum defect density slightly below the channel E_c (~0.2 eV below E_c), which can be exploited for improving gate stack reliability by introducing an interface dipole. The assumption is that the IL layer gives rise to such a dipole shifting the defect levels up with respect to the InGaAs E_c and shifting the minimum level into the device operating range. However, dipole engineering is not effective when a broad defect band without a clear minimum is present as is the case with Al_2O_3.

From the results shown till now, one can conclude that the 1 nm IL/3 nm HfO_2 stack shows an optimum for all electrical parameters although the D_{it} and N_{bt} levels are still above the targeted values (5 x 10^{11} /eV.cm^2 and 3 x 10^{10} /cm^2 respectively). However, when applying a PMA, we observe that the 1 nm IL/3 nm HfO_2 stacks degrades slightly i.e. a D_{it}, ΔN_{eff} increase and a reduction in the gamma value is observed in combination with a positive V_{fb} shift (Fig. 4-shaded bars). On the contrary, the stack with the Al_2O_3 cap does not degrade when a PMA is added. This result implies that although the initial gate stack properties are degraded when adding an Al_2O_3 cap, this cap can help in stabilizing the gate stack upon a PMA.

In conclusion, the IL approach provides an improved interface quality and increases the misalignment between the InGaAs channel E_f and the defect bands in the high-κ. The improved interface quality was proven to increase the performance of InGaAs MOS devices on both planar as well as nanowire architectures. However, the introduction of a cap layer is needed to provide a good thermal stability. Although Al_2O_3 can provide a higher stability for thermal processing, the unfavorable defect band distribution in the Al_2O_3 does not allow for dipole engineering. In the next part, we show that a $LaSiO_x$ cap can provide both an improved thermal stability and a more favorable defect band distribution.

3.2. Replacement of Al_2O_3 stabilization cap by $LaSiO_x$

Fig. 6 shows the multi-frequency C-V for various gate stacks in Table I and demonstrates the impact of the insertion of $LaSiO_x$ between the IL passivation and HfO_2. All the dot capacitors were measured after PMA at 400°C in forming gas anneal for 5 minutes. Although the CV's from the 1 nm IL/3 nm HfO_2 stack (Fig. 6A) and the 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 stack (Fig. 6C) look similar, one can notice that the frequency dispersion is further reduced when $LaSiO_x$ is inserted between the IL and HfO_2. As was discussed in the previous section, an Al_2O_3 cap degrades the gate stack properties slightly. Although the difference between the stacks with the IL are subtle, a large difference can be observed when the IL is removed (Fig. 6C'). These results imply that both the IL and the $LaSiO_x$ cap are needed to optimize the gate stack properties.

The D_{it}, extracted at mid gap, and capacitance equivalent thickness (CET) are shown in fig. 7. The insertion of the IL (A) reduces the mid gap D_{it} as was already reported in the previous section. The replacement of Al_2O_3 (B) by $LaSiO_x$ (C) further reduces the D_{it} and a reduction of the CET is also observed (CET=1.55 nm, EOT=1.15 nm). In addition, the total density of charging oxide traps drops to a value close to 1 x 10^{10} /cm^2 at E_{ox}

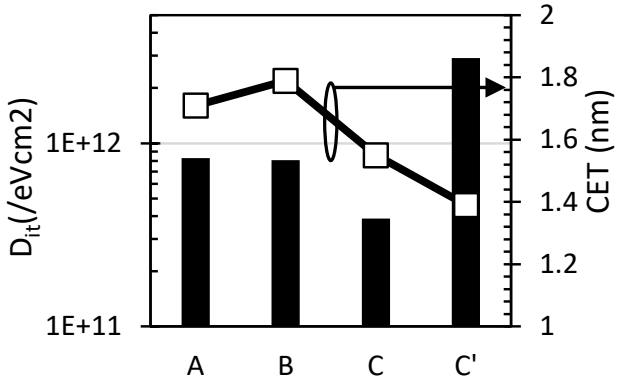


Figure 7: Mid gap D_{it} (bars) and CET (dots) extracted from the C-V data presented in fig. 6. Replacement of Al_2O_3 (B) by $LaSiO_x$ (C) reduces the D_{it} to less than 4 x 10^{11}/cm^{2}eV. The CET drops to 1.55 nm.

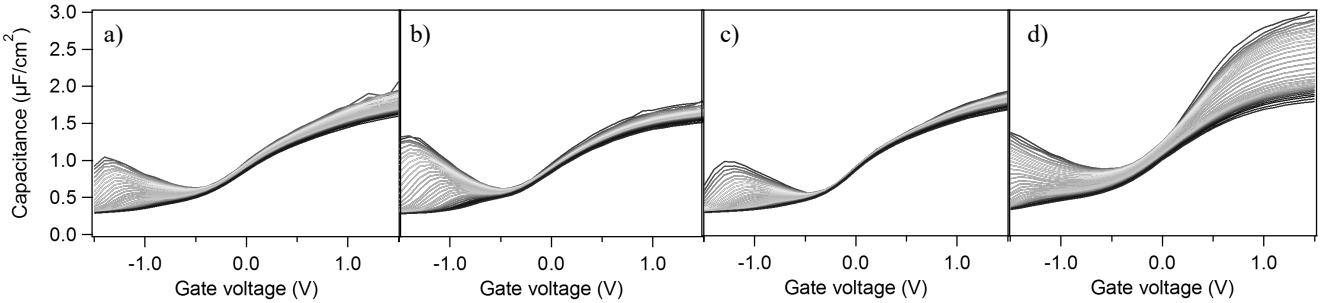


Figure 6. RT MF-CV's measured from dot capacitors of A) 1 nm IL/3 nm HfO_2, B) 1 nm IL/1 nm Al_2O_3/3 nm HfO_2, C) 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2, C') 1 nm $LaSiO_x$/3 nm HfO_2. Frequencies between 1 kHz-1 MHz (red-blue) were measured with logarithmic spacing. All capacitors received a PMA in forming gas for 5 minutes.

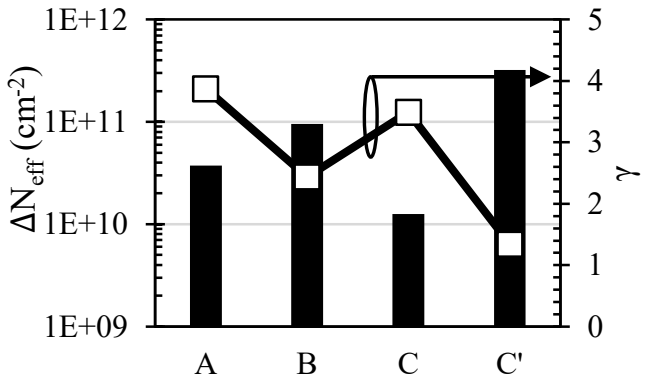


Figure 8: ΔN_{eff} (bars, left axis) extracted from C-V hysteresis (V_{start}=V_{fb} - 0.5 V) by converting the measured ΔV_{fb} into an effective charge sheet at the interface by: ΔN_{eff}=ΔV_{fb}*C_{ox}/q, and reported at charging equivalent field of 3.5 MV/cm [6]. Field acceleration exponent, γ, (dots, right axis) i.e. the slope of the log-log plot ΔN_{eff} vs oxide field which is a measure of the misalignment between channel carrier energy-i.e. InGaAs E_f and defect levels in the high-*k* [18].

= 3.5 MV/cm (Fig. 8) *meeting the target for 10 years operation* (10 years reliability criterion: ΔV_{fb} < 30 mV at 125°C. Assuming a typical time power law exponent of 0.13, the failure criterion projects ΔN_{eff} to less than 3 x 10^{10} /cm^2 at operating field, for an EOT of 1 nm). The reduction of oxide traps below the target level is the result of the combination of the IL layer and $LaSiO_x$ cap: while the IL layer increases the field dependence of the ΔN_{eff} (γ ~ 3.5), which is beneficial only for low voltage operation, the $LaSiO_x$ cap consistently reduces the density of charging defects irrespective of the operating voltage.

Fig. 9 shows the forward and backward sweep of gate voltage for the C-V hysteresis measured at different V_{max} for 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 stack. At low overdrive voltages, the hysteresis is negligible. Only at higher overdrive voltages, a clear hysteresis can be measured. Notice also that the starting voltage for these measurements was carefully chosen at V_{fb} - 0.5 V. As was reported in [25], the C-V hysteresis measurements can largely depend on the V_{start} condition, as is also the case for gate stack under evaluation (Fig. 10). The more negative V_{start} condition reflects the discharging of deep traps that can take place during OFF-state, when the device is biased deep into the depletion and is an equally important criterion to take into account. Measuring at V_{start}=V_{fb} - 0.5 V shows that oxide trap densities are within target. Starting from a more negative V_g, the oxide trap density increases and γ decreases. This observation indicates that this gate stack has a low

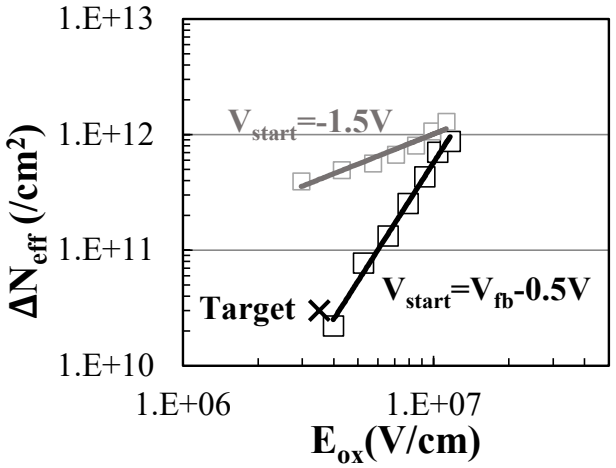


Figure 10: ΔN_{eff} as a function of applied E_{ox} for the stack with 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 for 2 different V_{start} conditions. V_{start}=-1.5V and V_{start}=V_{fb}-0.5V. The large dependence on the V_{start} conditions points to the presence of deep traps i.e. located at energies below InGaAs E_f at flatband.

amount of electron traps, but a more negative V_{start} charges hole traps, thereby increasing the total amount of traps during the measurement and decreasing the voltage dependence.

Therefore, BTI measurements were performed at both positive and negative stress conditions to capture the max V_{ov} for both stress conditions (Fig. 11). From these CV-BTI measurements, one can see that the 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 stack meets the

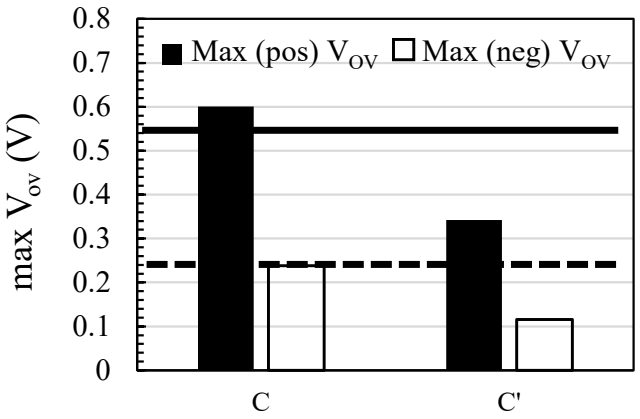


Figure 11: Max V_{ov} measured from positive and negative CV-BTI on MOS capacitors. For a target V_{DD}=0.8V, the max pos V_{ov} target (full line) is 0.53V and the max neg V_{ov} target (dashed line) is -0.26V (i.e. 1/3*V_{DD}).

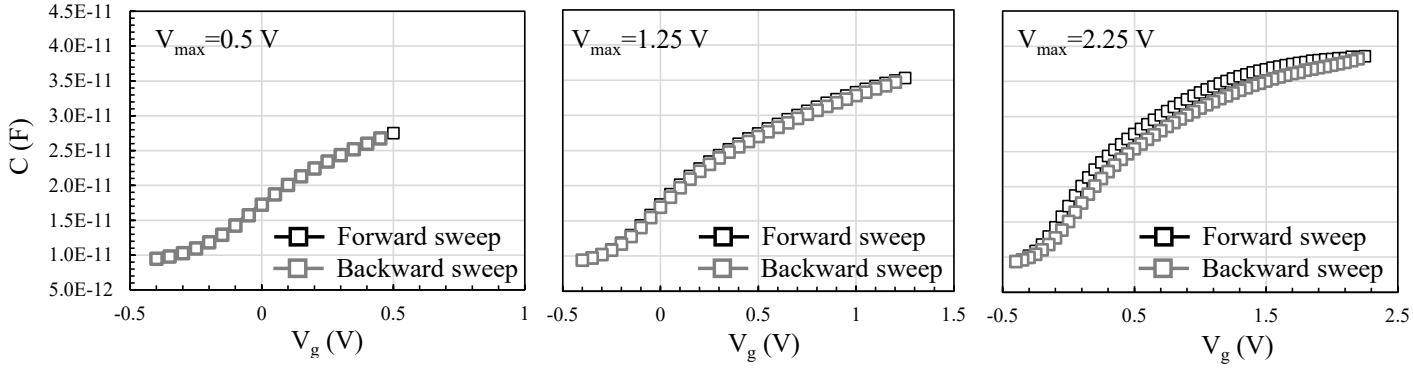


Figure 9. 100 kHz CV up and down traces of the 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 gate stack measured at 3 different V_{max}. V_{start} was V_{fb}-0.5V.

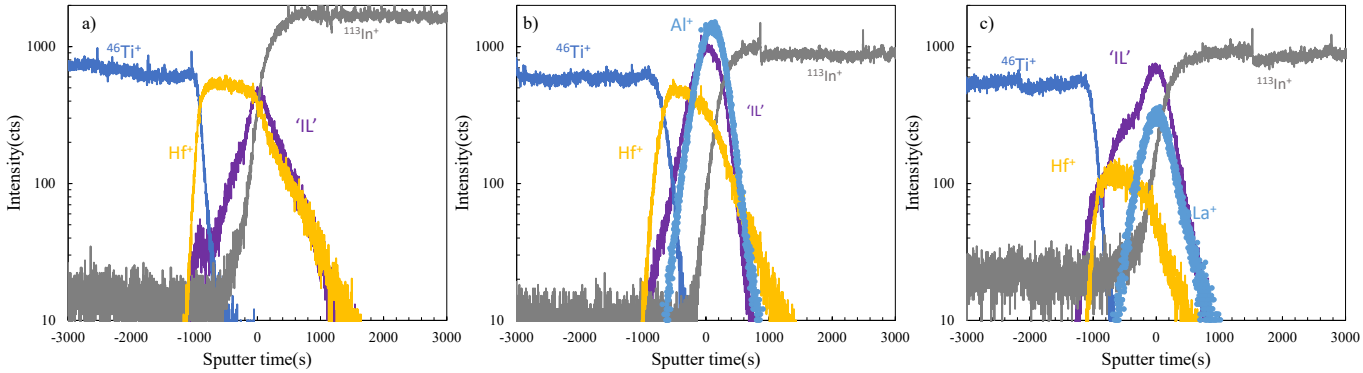


Figure 12. TOFSIMS positive ion profiles from MOSCAPs of stacks a) A, b) B and c) C from Table I. All layers are aligned at the peak of the IL atom profile.

max V_{ov} target (for a V_{DD}=0.8 V) for both stress conditions. When removing the IL layer, none of the targets are met. As III-V materials will be introduced in technologies operating at V_{DD}=0.5V and resulting target V_{ov}=0.33 V (=2/3*V_{DD}), this implies that these MOSCAP $In_{0.53}Ga_{0.47}As$ devices with the 1 nm IL/1 nm $LaSiO_x$/3 nm HfO_2 gate stack surpass the IIIV reliability requirements.

Stabilization of the IL by the Al_2O_3 or $LaSiO_x$ cap is confirmed from the TOFSIMS profiles measured on dot capacitors that received a PMA (Fig. 12**)**. These profiles show the mixing of the Al atoms (from the Al_2O_3) or La atoms (from the $LaSiO_x$) with the IL. Diffusion of the IL atoms down into the InGaAs substrate is clearly present if no cap layer is deposited between IL and the HfO_2 (Fig. 12a). When an Al_2O_3 or $LaSiO_x$ cap is added on top of the IL (Figs. 12b and 12c) , a steep slope of the IL elements down into the substrate is observed. The hypothesis is that the mixing of the Al or La atoms with the IL atoms is increasing the stability of the IL. Although the chemistry from both Al_2O_3 and $LaSiO_x$ cap layers can provide stabilization of the interaction of the IL with the substrate, the electrical characteristics of the $LaSiO_x$ cap are more favorable.

Modeling of the shallow and deep defect bands using the approach as described in [26], was done and explains the effect of the $LaSiO_x$ cap **(**Fig. 13**)**. The introduction of the IL layer narrows the distribution of the shallow defect band as well as shifts the mean energy of that band to shallower energies. Adding $LaSiO_x$ in between the IL and HfO_2 further reduces the charging defect density. This reduction can be attributed to the stabilization that is provided by the $LaSiO_x$ cap.

3.3. MOSFET device characteristics

Figs. 14 and 15 show the I_{ds}-V_{gs} and g_m-V_{gs} curves of the devices fabricated with stacks B (Al_2O_3 cap) and C ($LaSiO_x$ cap). Excellent device characteristics in the form of high I_{on}/I_{off} ratio (> $2x10^6$, > $8x10^5$ at V_{ds}=0.05 V and 0.5 V respectively) and low SS_{lin} (< 75 mV/dec) were extracted for both the gate stacks. The device characteristics for stacks B, C, C', and C'' are given in Fig. 16 and 17. The on-state current increases (Fig. 16a and b**)** and the SS_{lin} decreases (Fig. 17) when Al_2O_3 (B) is replaced by $LaSiO_x$ (C) between the IL layer and HfO_2. When scaling the HfO_2 to 2 nm (C''), a CET of 1.46 nm (EOT of 1.06 nm) is achieved with a

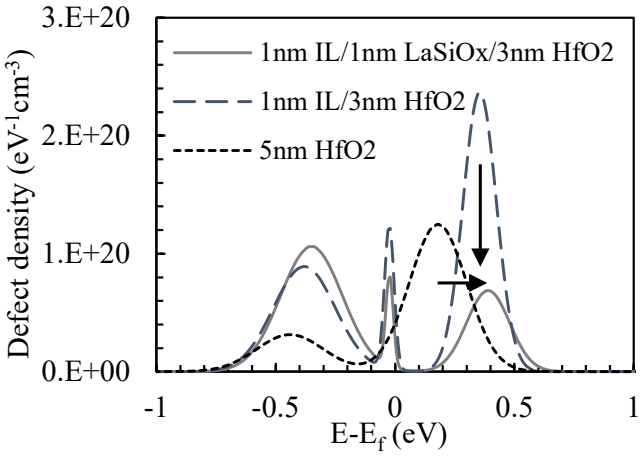


Figure 13**:** Defect density profile obtained by modeling the ΔV_{th} vs V_{gate} as described in [25]. Insertion of the IL layer causes narrowing of the shallow defect band, thus enhancing γ and shifting the mean defect level of this band to shallower energy [19] but increases the density of defects at deeper energies. Insertion of $LaSiO_x$ consistently reduces the total defect density (both shallow and deep energies).

further improvement in the on-characteristics, and with a reduction in the SS_{lin} to 68 mV/dec. Note that the stack without the IL layer (1 nm $LaSiO_x$/3 nm HfO_2 stack only, C'), exhibits high on-current compared to the stack with the 1 nm IL/1 nm Al_2O_3/3 nm HfO_2 but

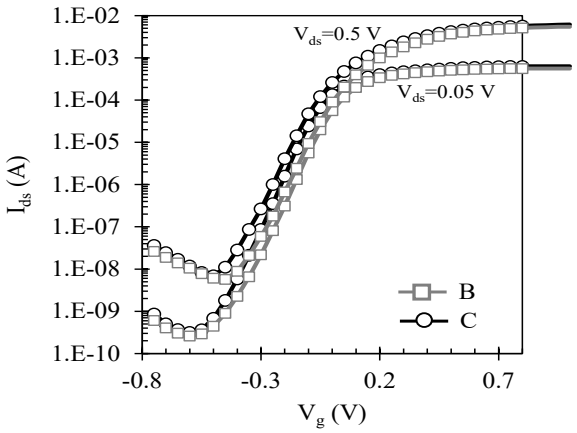


Figure 14**:** I_{ds}-V_{gs} for L_g=10 µm, W=100 µm device.

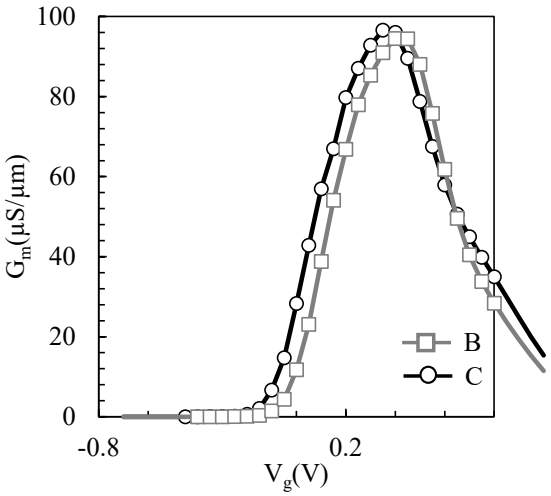


Figure 15: G_m-V_{gs} for L_g=10µm, W=100µm device.

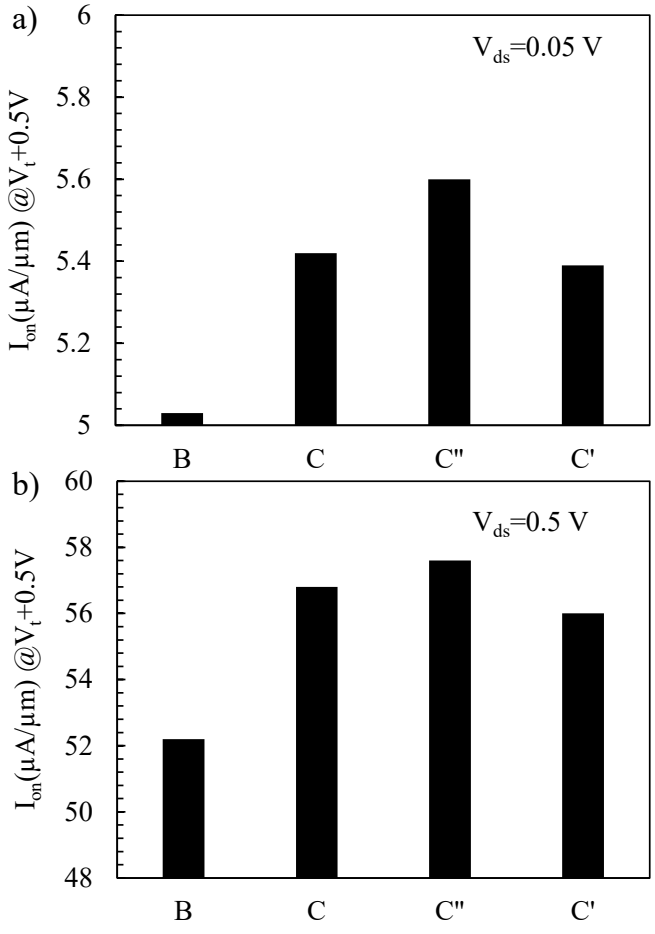


Figure 16: I_{on} extracted at V_t+0.5V and for V_{ds}=0.05V (a) and V_{ds}=0.5V (b), W of the devices = 100µm, Lg=10µm. Replacement of Al_2O_3 (B) by $LaSiO_x$ (C) in the gate stack improves I_{on}. Further improvement can be seen when the gate stack is scaled down (C'').

also a high SS_{lin} reflecting a higher mid gap D_{it} extracted from the MOS capacitor C-V's (Fig. 7 and 8). Only slight changes in V_{th_lin} (~50 mV) are observed when changing the stabilization cap (Fig. 18).

The mobility data (Fig. 19) reveal that while the dominant

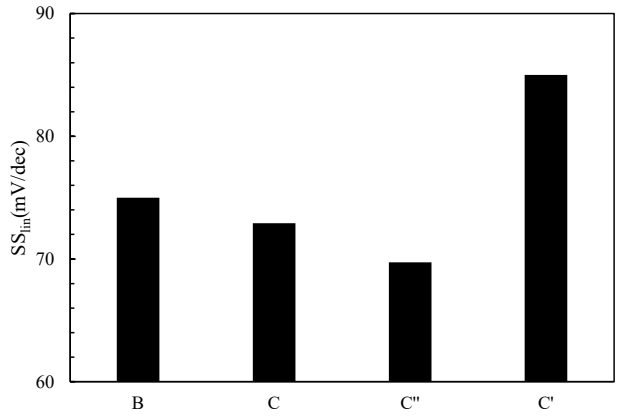


Fig. 17 SS_{lin} for various gate stacks. Lg=50 µm. SS_{lin} reduces by replacing Al_2O_3 (B) with $LaSiO_x$ (C) in the gate stack. Scaling the HfO_2 (C'') thickness leads to further reduction.

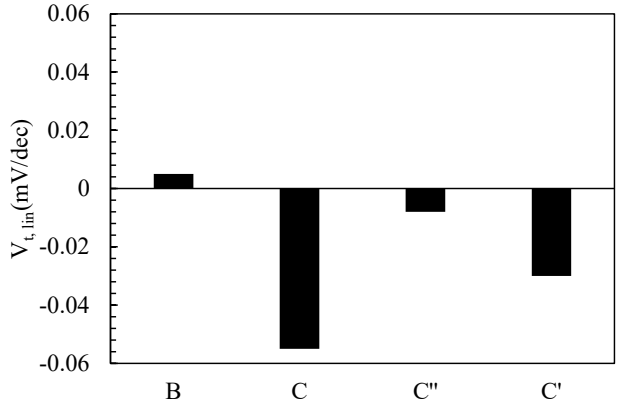


Figure 18 $V_{t,lin}$ for various gate stacks. L_g=50 µm. Slight Vt_{lin} reduction by replacing Al_2O_3 (B) with $LaSiO_x$ (C) in the gate stack is observed.

impact of insertion of the IL layer is on the peak mobility (μ_{eff} at peak = 3531 cm^2/V.s), the replacement of the Al_2O_3 by $LaSiO_x$ increases the electron mobility at high charge carrier density.

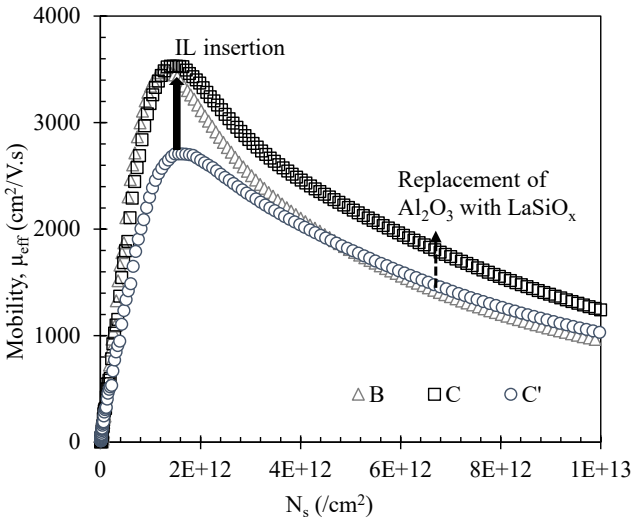


Figure 19 Improvement in the mobility at low N_s (due to the introduction of IL) and high N_s (as a result of replacement of Al_2O_3 with $LaSiO_x$). Mobility was extracted from L_g=50 µm devices.

This observation suggests a reduction in the roughness using $LaSiO_x$ in the gate stack instead of Al_2O_3. In addition, BTI measurements on MOSFETs confirm the good reliability observed on simple MOS capacitor structures. ΔN_{eff} and γ values were extracted under both positive and negative stress conditions (Fig. 20). Under both stress conditions, we can observe a reduction of ΔN_{eff} when adding the IL layer and further reduction is achieved by insertion of the $LaSiO_x$ cap. In parallel, the value of γ increases when adding the IL layer but levels off when inserting the $LaSiO_x$. This confirms both the

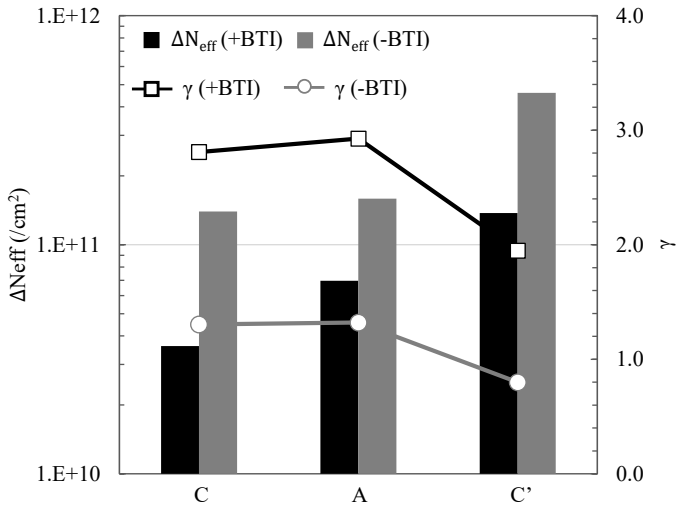


Figure 20 ΔN_{eff} and γ extracted from BTI measurements on MOSFETs applying both positive and negative stress conditions. The data from MOSFETs confirm the trends observed on MOSCAPs.

observation in measurement data and modeling results from the MOSCAP data i.e. the IL layer provides the misalignment between the defect levels in the high-κ and the InGaAs E_f while the $LaSiO_x$ reduced the total charging oxide defects. At negative stress conditions, the differences are less pronounced.

The extrapolated max V_{ov} and V_{un} at 10 years are shown in Fig. 21. Benchmarking the max V_{ov} of the newly proposed gate stack against Si gate stacks (Fig. 22) shows that by using the optimized gate stack III-V devices can meet even the more stringent reliability targets used for a scaled Si technology.

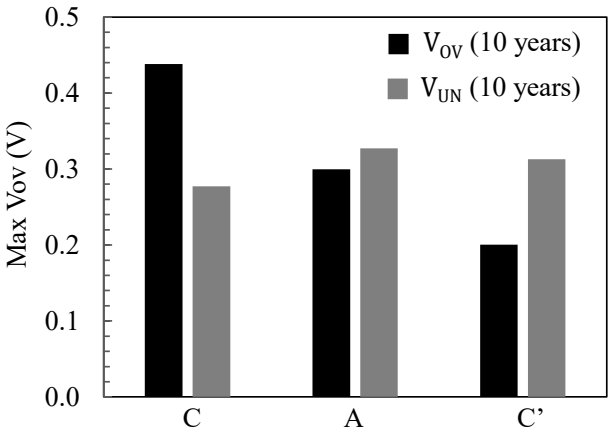


Figure 21. Max V_{ov} and V_{un} from BTI measurements on MOSFETs with different gate stacks.

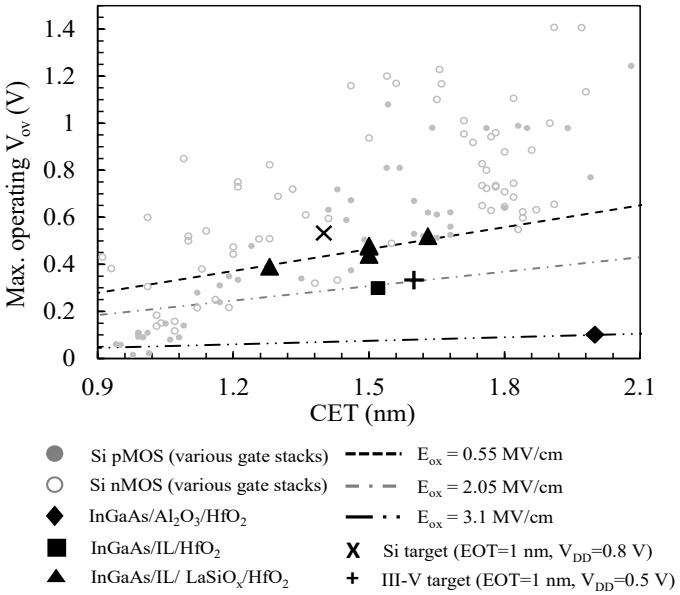


Figure 22. Max. operating overdrive voltage as a function of CET for various gate stacks on Si (p- and n-MOS) and InGaAs. The devices with the IL/$LaSiO_x$/HfO_2 gate stack surpass the relability target for III-V semiconductor based technology.

4. Conclusions

We have shown in this work the improved interface and reliability properties of the IL layer. A D_{it} and ΔN_{eff} reduction are accompanied with a negative V_{fb} shift and an increase in the value of γ. This suggests the formation of an interface dipole. When a favorable defect band distribution is present as is the case for HfO_2, dipole engineering can be applied to misalign the defect band with the InGaAs channel E_f. However, the thermal stability of this stack is limited and a severe intermixing of the IL layer with the InGaAs substrate at typical thermal budget of a IIIV gate stack, degrades the initial improvement seen from the IL layer.

A cap is needed to stabilize this interaction and both Al_2O_3 and $LaSiO_x$ can act as a stabilization layer by La or Al mixing with the IL-layer atoms. While the physical properties are very similar, the electrical properties in terms of oxide traps are very different. The different behavior can be explained by the presence of a broad defect band in Al_2O_3 making this cap layer not suitable for dipole engineering. On the other hand, $LaSiO_x$ shows a more favorable defect band distribution and a further reduction of the oxide defects is observed. In addition, the interface properties are improved and low CET (~1.5 nm) was maintained.

Excellent device characteristics were demonstrated combining the IL passivation layer with the bi-layer stack consisting of 1 nm $LaSiO_x$/3 nm HfO_2. III-V devices with record mobility and record reliability performance at scaled EOT were demonstrated. In addition, we show that this stack can be further scaled down to 1nm EOT without loss of performance or reliability.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Sioncke, J. Franco, A. Vais, V. Putcha, L. Nyns, A. Sibaja-Hernandez, R. Rooyackers, S. Calderon Ardila, V. Spampinato, A. Franquet, J.W. Maes,

Q.Xie, M.Givens, F. Tang, X. Jiang, M. Heyns, D. Linten, J. Mitard, A. Thean, D. Mocuta and N. Collaert, "First demonstration of ~3500 cm^2/V-s electron mobility and sufficient BTI reliability (max V_{ov} up to 0.6 V) $In_{0.53}Ga_{0.47}As$ nFET using an IL/$LaSiO_x$/HfO_2 gate stack", in IEEE Symposium on VLSI Technology 2017, pp. T38-T39, 2017. https://doi.org/10.23919/VLSIT.2017.7998192

[2] A. Vais, A. Alian, L. Nyns, J. Franco, S. Sioncke, V. Putcha H. Yu, Y. Mols, R. Rooyackers, D. Lin, J.W. Maes, Q. Xie, M. Givens, F. Tang, X. Jiang, A. Mocuta, N. Collaert, K. De Meyer, A. Thean, "Record mobility (μ_{eff}~ 3100 cm^2/Vs) and reliability performance (V_{ov}~ 0.5 V for 10yr operation) of $In_{0.53}Ga_{0.47}As$ MOS devices using improved surface preparation and a novel interfacial layer" in IEEE Symposium on VLSI Technology, pp. 140-141. 2016. https://doi.org/10.1109/VLSIT.2016.7573410

[3] C. B. Zota, F. Lindelöw, L.-E. Wernersson, and E. Lind. "InGaAs nanowire MOSFETs with I_{ON}= 555 μA/μm at I_{OFF}= 100 nA/μm and V_{DD}= 0.5 V" in IEEE Symposium on VLSI Technology, pp. 1-2. IEEE, 2016. https://doi.org/10.1109/VLSIT.2016.7573418

[4] N. Waldron, S. Sioncke, J. Franco, L. Nyns, A. Vais, X. Zhou, H. C. Lin, G. Boccardi, J. W. Maes, Q. Xie and M. Givens, F. Tang , X. Jiang , E. Chiu, A. Opdebeeck, C. Merckling, F. Sebaai, D. van Dorp, L. Teugels, A. Sibaja Hernandez, K. De Meyer, K. Barla, N. Collaert, A. Thean, "Gate-all-around InGaAs nanowire FETS with peak transconductance of 2200 μS/μm at 50 nm L_g using a replacement Fin RMG flow" in IEEE International Electron Devices Meeting (IEDM), pp. 31-1. IEEE, 2015. https://doi.org/10.1109/IEDM.2015.7409805

[5] R. Suzuki, N. Taoka, M. Yokoyama, S. Lee, S. H. Kim, T. Hoshii, T. Yasuda, W. Jevasuwan, T. Maeda, O. Ichikawa, N. Fukuhara, M. Hata, M. Takenaka, and S. Takagi, " 1-nm-capacitance-equivalent-thickness HfO_2/Al_2O_3/InGaAs metal-oxide-semiconductor structure with low interface trap density and low gate leakage current density," Appl. Phys. Lett. **100**(13), 132906-1–132906-3 (2012). https://doi.org/10.1063/1.3698095.

[6] F Palumbo, M Eizenberg, "Degradation characteristics of metal/Al_2O_3/n-InGaAs capacitors", Journal of Applied Physics, 115 (1), 014106 (2014). DOI: https://doi.org/10.1063/1.4861033

[7] G. Jiao, C. Yao, Y. Xuan, D. Huang, P.D. Ye, and M.-F. Li, "Experimental Investigation of Border Trap Generation in InGaAs nMOSFETs With Al2O3 Gate Dielectric Under PBTI Stress", IEEE Trans. Electron Dev., 59, pp. 1661-67 (2012). https://doi.org/10.1109/TED.2012.2190417

[8] I Krylov, B Pokroy, M Eizenberg, and D Ritter, "A comparison between HfO_2/Al_2O_3 nano-laminates and ternary Hf_xAl_yO compound as the dielectric material in InGaAs based metal-oxide-semiconductor (MOS) capacitors", Journal of Applied Physics, 120 (12), 124505 (2016). https://doi.org/10.1063/1.4962855

[9] K. Tang, R. Droopad, and P. C. McIntyre. "Border Trap Density in Al_2O_3/InGaAs MOS: Dependence on Hydrogen Passivation and Bias Temperature Stress", ECS Trans., 69 (5), pp. 53-60 (2015). https://doi.org/10.1149/06905.0053ecst

[10] C. L. Hinkle, M. Milojevic, B. Brennan, A. M. Sonnet, F. S. Aguirre-Tostado, G. J. Hughes, E. M. Vogel, R. M. Wallace, "Detection of Ga suboxides and their impact on III-V passivation and Fermi-level pinning", Appl. Phys. Lett. 94, 162101 (2009). https://doi.org/10.1063/1.3120546

[11] D. Lin, A. Alian, S. Gupta, B. Yang, E. Bury, S. Sioncke, R. Degraeve, M. L. Toledano, R. Krom, P. Favia, H. Bender, M. Caymax, K. C. Saraswat, N. Collaert, A. Thean, "Beyond interface: The impact of oxide border traps on InGaAs and Ge n-MOSFETs", in Proceedings of IEEE International Electron Device Meeting Technical Digest, 28.3.1 (2012). https://doi.org/10.1109/IEDM.2012.6479121

[12] S. Deora, G. Bersuker, W.-Y. Loh, D. Veksler, K. Matthews, T. W. Kim, R. T. P. Lee, R. J. W. Hill, D.-H. Kim, W.-E. Wang, C. Hobbs, P. D. Kirsch, "Positive Bias Instability and Recovery in InGaAs Channel nMOSFETs", IEEE Transactions on Device and Materials Reliability, 13(4), 507 (2013). https://doi.org/10.1109/TDMR.2013.2284376

[13] J. Franco A. Alian, B. Kaczer, D. Lin, T. Ivanov, A. Pourghaderi, K. Martens, Y. Mols, D. Zhou, N. Waldron, S. Sioncke, T. Kauerauf, N. Collaert, A. Thean, M. Heyns, and G. Groeseneken, "Suitability of high-k gate oxides for III-V devices: a PBTI study in $In_{0.53}Ga_{0.47}As$ devices with Al_2O_3", in Proceedings of IEEE International Reliability Physics Symposium Technical Digest, 6A.2.1 (2014). https://doi.org/ 10.1109/IRPS.2014.6861098

[14] N. Conrad, M. Si, S. H. Shin, J. J. Gu, J. Zhang, M. A. Alam, P. D. Ye, "Low-frequency noise and RTN on near-ballistic III–V GAA nanowire MOSFETs", in Proceedings of IEEE International Electron Devices Meeting, 20.1.1 (2014). https://doi.org/10.1109/IEDM.2014.7047086

[15] N. Taoka, M. Yokoyama, S. H. Kim, R. Suzuki, S. Lee, R. Iida, T. Hoshii, W. Jevasuwan, T. Maeda, T. Yasuda, O. Ichikawa, N. Fukuhara, M. Hata, M.Takenaka and S. Takagi, "Impact of Fermi Level Pinning Due to Interface Traps Inside the Conduction Band on the Inversion-Layer Mobility in InGaAs Metal–Oxide–Semiconductor Field Effect Transistors". IEEE Trans. Dev. Mat. Rel. 13, 456 (2013). https://doi.org/10.1109/TDMR.2013.2289330.

[16] S. Stemmer, V. Chobpattana, and S. Rajan, "Frequency dispersion in III-V metal-oxide-semiconductor capacitors", Appl. Phys. Lett. 100, 233510 (2012). https://doi.org/10.1063/1.4724330.

[17] S. . Johansson, M. Berg, K. M. Persson and E. Lind, "Transconductance method for characterization of border traps", *IEEE Trans. Electron Devices*, vol. 60, no. 2, February 2013. https://doi.org/ 10.1109/TED.2012.2231867

[18] J. Franco, B. Kaczer, Ph. J. Roussel, J. Mitard, S. Sioncke, L. Witters, H. Mertens, T. Grasser, and G. Groeseneken, "Understanding the suppressed charge trapping in relaxed-and strained-Ge/SiO_2/HfO_2 p-MOSFETs and implications for the screening of alternative high-mobility substrate/dielectric CMOS gate stacks" in Proceedings of IEEE International Electron Devices Meeting, 15.2.1 (2013). https://doi.org/ 10.1109/IEDM.2013.6724634

[19] J. Franco, A. Vais, S. Sioncke, V. Putcha, B. Kaczer, B-S. Shie, X. Shi, R. Mahlouji, L. Nyns, D. Zhou, and N. Waldron, "Demonstration of an InGaAs gate stack with sufficient PBTI reliability by thermal budget optimization, nitridation, high-k material choice, and interface dipole" in 2016 IEEE Symposium on VLSI Technology, , pp. 1-2., 2016. https://doi.org/ 10.1109/VLSIT.2016.7573371

[20] A. Alian, M. A. Pourghaderi, Y. Mols, M. Cantoro, T. Ivanov, N. Collaert, and A. Thean, "Impact of the channel thickness on the performance of ultrathin InGaAs channel MOSFET devices", in Proceedings of IEEE International Electron Device Meeting, pp.16.6.1-16.6.4, 2013. https://doi.org/ 10.1109/IEDM.2013.6724644

[21] E. H. Nicolian and J. R. Brews, MOS Phys. and Tech., p. 215-216, Wiley Interscience, New York (2003).

[22] R. Winter, J. Ahn, P. C. McIntyre, and M. Eizenberg, "New method for determining flat-band voltage in high mobility semiconductors", Journal of Vacuum Science & Technology B **31**, 030604 (2013). https://doi.org/ 10.1116/1.4802478

[23] D. K. Schroder, Semiconductor material and device characterization. John Wiley and Sons, 2006, ISBN: 978-0-471-73906-7.

[24] B. Kaczer, T. Grasser, P.J. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, and G. Groeseneken, "Ubiquitous relaxation in BTI stressing—New evaluation and insights" in proceedings of IEEE International Reliability Physics Symposium, 2008, pp. 20-27, 2008. https://doi.org/ 10.1109/RELPHY.2008.4558858

[25] A. Vais, J. Franco, H.-C. Lin, N. Collaert, A. Mocuta, K. De Meyer, and A. Thean, "Impact of starting measurement voltage relative to flat-band voltage position on the capacitance-voltage hysteresis and on the defect characterization of InGaAs/high-k metal-oxide-semiconductor stacks", Appl. Phys. Lett. 107, 223504 (2015). https://doi.org/ 10.1063/1.4936991

[26] V. Putcha, J. Franco, A. Vais, B. Kaczer, S. Sioncke, D. Linten and G. Groeseneken, "Impact of slow and fast oxide traps on $In_{0.53}Ga_{0.47}As$ device operation studied using CET maps", in proceedings of IEEE International Reliability Physics Symposium 2017, XT-8.1 (2017). https://doi.org/ 10.1109/IRPS.2018.8353603

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 3, 43-50 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

A Cyber-Vigilance System for Anti-Terrorist Drives Based on an Unmanned Aerial Vehicular Networking Signal Jammer for Specific Territorial Security

Dhiman Chowdhury[*,1], Mrinmoy Sarkar[2], Mohammad Zakaria Haider[3]

[1]*Electrical Engineering, University of South Carolina, SC 29208, U.S.A.*

[2]*Electrical and Computer Engineering, North Carolina A & T State University, NC- 27411, U.S.A.*

[3]*Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh*

ARTICLEINFO

Article history:
Received: 11 April, 2018
Accepted: 02 May, 2018
Online: 20 May, 2018

Keywords:
Anti-terrorist drive
Band-pass filter
Cyber-vigilance system
Motion control
Networking signal jammer
Territorial security
Unmanned aerial vehicle
Video surveillance

ABSTRACT

During sudden anti-terrorist drives conducted by the law enforcement agencies, a localized cyber security system happens to be a special tactic to avert the unprecedented massacre and gruesome fatalities against the residents of that area by disconnecting the affected territory from the rest of the world; so that the militants and their outside accomplices cannot communicate with each other and also the terrorists cannot go through the ongoing apprehensive operation via wireless communications. This paper presents a novel framework of an unmanned aerial vehicular networking signal jammer which is oriented to block incoming and outgoing signals of all frequencies transmitted from a specifically marginalized territory scanned and explored by the aerial vehicle. During such a cyber-vigilance operation, the aerial vehicle is equipped with a transmitter and an auto-tuning band-pass filter module with automatic regulation of center frequencies according to the surrounding networking signals, which are considered to be the suppressing noise parameters. In order to restrict the signal blocking operation within the militant hub, the aerial vehicle with the network terminator is controlled to navigate within a particular boundary of a residential area and its navigation is continuously mapped and stored for effective evacuation process directed to save the innocent stranded people. A very low frequency (VLF) metal detector has been designed to trace the explosives and buried landmines inside the exploration arena. An algorithm for 3-D mapping of the metal traces detected by the aerial navigator has been presented in this paper. Signal blocking, metal tracing and stable confined movements have been tested where the testbed is provided with signals of different frequencies along with variation in dimensions of the testing region to evaluate the reliability of the proposed framework.

1 Introduction

In order to commit crimes with ease and for ensuring their maximum safety with encrypted identity from the law enforcement personnel, the militants choose places full of common people where the abodes are suburbs or localities with a dense population and residential housing structures. During certain anti-militant drives, the first and foremost security aspect is to evacuate the territory so that the innocent victims seized in the attacked area remain unharmed and minimum household materials are to be damaged. Due to the exposure of Internet and wireless communication provision, the militants supposedly utilize networking signals and Internet facilities to penetrate the comprehensive anti-terrorist operation and they often materialize communication with their outside accomplices to assist them to flee from the drive. If the terrorists

[*]University of South Carolina, SC 29208, Columbia, U.S.A., +19712050907 & dhiman@email.sc.edu

https://dx.doi.org/10.25046/aj030306

become successful to pervade the vigilance system, they can cause unprecedented human fatalities of the stranded local people.

This paper is an extended version of the work reported in [1]. An unmanned aerial vehicle equipped with a networking signal jammer has been implemented for monitoring and controlling an anti-militant mission. The aerial vehicle is manually and semi-autonomously operated to explore within a specified region of interest with an on-board camera providing consistent video feedback and an electronic device subsuming an auto-tuning band-pass filter with a transmitter and a segregated power module. The extensions in this paper are the inclusion of a consistent 3-D mapping of the detected metallic substances, performance evaluation of the signal blocking operation in 4G networks, test-bed analysis for accurate metal tracings and deviations in 3-D navigations of the employed aerial vehicle within certain territorial regions.

The employed prototype and its motion planning algorithms are reported in [2] where to restrict the aerial exploration inside a marginalized trajectory, the path planning algorithm has been modified. The process of the autonomous concatenated map generation of the traversed region is described in [2] and [3]. The auto center frequency reconfigurable bandpass filter to substantiate networking signal jammer is based on a novel design proposed in [4] where an ultra high frequency and high quality factor integrated Gm-C technology based bandpass filter is presented which is fabricated using the modified Nauta's Transconductor principle.

Traditionally, a signal jammer enacts the role of an adaptive source of interfering noise suppressed onto a propagating message signal at the node of eavesdroppers so that the confidentiality of the relevant wireless protocol is maintained by not allowing any outsider to decode the inherent messaging information. In the papers [5]-[8] the degrading security concerns in military operations based on wireless communication, jamming based physical layer security prospects and particular environmental reliability challenges incorporated in sensor networks are reported respectively. The paper presented in [8] documents an innovative anti-deception jamming method for radar sensor networks whereas in [9] a detailed perspective analysis of the secrecy transmission capacity of hostile jammers is articulated. The paper presented in [10] describes a novel outage-based characterization approach for joint relay and jammer selection scheme in case of channel state information feedback delays.

To enhance the strength of the signal of interest in a wireless communication scheme, jamming attacks should be effectively retaliated. In the works reported in [11] and [12] the investigating methodologies related to a MIMO interference cancellation based jamming resilient system, an energy harvesting jammer to minimize secrecy outage for powered communication networks and optimal stopping theory dependent jammer selection process for secured radio networks are documented respectively.

The proposed aerial vehicular signal jamming system incorporates auto tuning band-pass filter module with consistent video surveillance and automatic map generation capabilities. The vehicle is boarded with a very low frequency discriminative metal detector circuit to trace different types of explosive materials. The locomotion of the aerial route is confined to a specified territory with the provision of manual and semi-autonomous motion control algorithms. The implemented system with its salient features, applied signal jamming technique, motion control and mapping algorithms along with explosives detection technique and test-bed results are articulated in this paper section-by-section.

2 Implemented Cyber-Vigilance System and Proposed Salient Features

The tactical security system based on unmanned aerial vehicle facilitated with networking signal jammer is solicited to be deployed in case of anti-militant and anti-terrorist drives conducted by the law enforcement agencies. The investigated methodologies, salient features and entities, control algorithms and constituents of the developed framework are articulated in the following excerpts.

2.1 Unmanned Aerial Vehicle

The detailed specifications of the aerial vehicle are described in [2] and Fig. 1 shows the demonstrated aircraft. Fig. 2 presents the vehicle at an elevation. As mentioned in [2], Q450 V3 is the main quad frame with glass fiber platform with arms constructed of polyamide nylon. There are four 820 kVA BLDC motors and for solemnizing motion and elevation control there are two normal and two pusher propellers. Being exposed to a payload of 0.5-0.55 kg and a single shot flight duration of approximately 14 minutes, the entire aerial structure weighs around 1.4 kg.

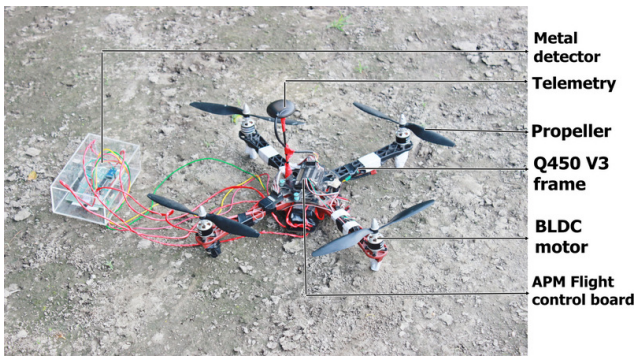


Figure 1: Employed unmanned aerial vehicle [the background is a muddy surface of the test arena]



Figure 2: Vehicle in flight mode

An open source ground station application for MAV communication, called APM Mission Planner 2.0, is utilized to control the flight modes, longitude and altitude parameters of the vehicle. Mission Planner is interfaced with manual and semi-autonomous motion planning algorithms developed on C language platform with continuous data logging with the GPS waypoints and control events. As presented in [2], there is a simulation interface of the flight control phenomena and the associated tuning parameters.

2.2 Networking Signal Jammer

To block the surrounding networking signals propagating in the terrorized region, a novel band-pass filter module along with a radio transmitter is utilized as the signal jammer. To self-calibrate the jamming process, the radio transmitter is housed with the band-pass filter and the overall jammer set-up is attached to the surface of the vehicle.
According to the design considerations and technology proposed in [4], TSMC 0.18 μm CMOS fabrication process based Gm-C band-pass filter module is used where active transconductor components and six CMOS inverters are incorporated. The quality factor is maintained more than 40 and the adjustable center frequency of the operating bandwidth can be configured within the range of 797.4 MHz to 819.5 MHz. The automatic tuning procedure is enabled following an analog Phase-Locked-Loop (PLL) technique based on an adaptive integrated Voltage-Controlled-Oscillator (VCO). As described in [4], an inherently symmetric improved Nauta's Transconductor building block is introduced.
The improved Nauta's Transconductor configuration, device parameters, equivalent circuit of the fabricated Gm-C band-pass filter, analog PLL topology for automatic frequency tuning system, architecture of the adaptive VCO and relative performance evaluation are presented in [4]. The signal jammer-cum-adaptive band-pass filter module tunes its center frequencies within an approximate range of 797 MHz to 819 MHz with a unity gain and it blocks other frequency signals as the commensurate suppressing noise parameters. The data acquisition rate of the Wi-Fi camera is constantly kept at 800 MHz and hence the video surveillance is not interrupted during the entire systematic operation.
Several band-pass filter topologies for communication media and transmission channels with low-loss, cost-reliability and high selectivity factors have been investigated like the work presented in [13]. This paper documents a microstrip band-pass filter with electronically tunable notch response which keeps the bandwidth within a range of 2.9 GHz to 5.8 GHz. The operating bandwidth range lies in the interest communication signal transmission range and hence this approach is not employed in this proposed vigilance system. A novel design of frequency reconfigurable transmission antennas with the provision of automatic switching features is proposed in [14] where a voltage doubler circuit is utilized to convert a Radio Frequency value to its equivalent DC reference value. In [15], a FPGA based FIR band-pass filter is proposed for satellite communication and in [16], a spectral-parameter-approximation based variable digital filter with adjustable center frequencies is described. Another sophisticated technology for developing tunable band-pass filter is presented in [17] where an integrated coupled resonator optical waveguide based band-pass filter is explained for photonic signal processing. The methodologies and implemented band-pass filters in [15],[16] and [17] are not suitable for practical implementation for an aerial project because of the design complexity and very high cost of deployment.

2.3 Motion Control Algorithm of the Aerial Vehicle

Aerial navigation of the explorer is subject to control in order to ensure a reliable and effective surveillance operation. The aerodynamics enact as a key factor to determine the degrees of freedom of the aerial movement of the vehicle. Air friction and vibration of the mechanical structure are two considerable aspects in case of motion control application.
The aerial exploration of the quadrotor is consistently monitored and controlled from a base station. Initially the uplifting and elevation of the vehicle are controlled manually and then a modified Point-to-Point (PTP) path planning algorithm, as described in [2], controls the aerial movement semi-autonomously. The PTP algorithm is very much similar to the process introduced in [18] where the monitoring arena is divided into several cells of uniform dimensions.
In [19] and [20], path planning algorithms of unmanned aerial vehicles offering coverage missions and geo-fencing activities are proposed respectively where energy and maneuverability constraints and multi-objective optimization principle based on a multi-gene structure are considered. Genetic algorithm formulates multi-point search space and an improved path control method is proposed in [21] which is a more sophisticated version of the customized PTP algorithm. In [22], a time optimal two-dimensional path planning

approach is documented along with obstacle avoidance provision. Several sophisticated motion control techniques for unmanned aerial vehicles have been articulated in [23], [24] and [25].
The basic control algorithm to decide the navigation trajectory of the demonstrated aerial vehicle follows the one presented in [2] with a modification of insertion of three dimensional movement within a bounded arena with limited specifications. The aerial vehicle is subject to scan the maximum coverage area of the region of interest by moving from a point to its nearest point of exploration. The customized algorithm for the aerial motion control is presented in Algorithm 1.
Apart form that GPS coordinates and orientation sensors embedded onto the vehicular structure enables automatic mapping of the stored position values of the navigating and continuously moving quadrotor.

Algorithm 1 Algorithm for Motion Control

Array of X-axis coordinates of center of each grid $\rightarrow X$
Array of Y-axis coordinates of center of each grid $\rightarrow Y$
Array of Z-axis coordinates of center of each grid $\rightarrow Z$
Dummy variables $\rightarrow i, j, k$
Maximum index of X-array $\rightarrow I$
Maximum index of Y-array $\rightarrow J$
Maximum index of Z-array $\rightarrow K$
Coordinates of the present grid $\rightarrow x, y, z$
INITIALIZE:
$i \leftarrow 0$
$j \leftarrow 0$
$k \leftarrow 0$
START:
While $(j < J)$
$x \leftarrow x[i]$
$y \leftarrow Y[j]$
$y \leftarrow Z[k]$
if $j\%2 = 0$ **then**
  $i \leftarrow i + 1$
  $k \leftarrow k + 1$
else
  $i \leftarrow i - 1$
  $k \leftarrow k - 1$
end if
if $i = I$ and $k = K$ **then**
  $i \leftarrow i - 1$
  $k \leftarrow k - 1$
  $j \leftarrow j + 1$
  if $i = -1$ and $k = -1$ **then**
    $i \leftarrow i + 1$
    $k \leftarrow k + 1$
    $j \leftarrow j + 1$
  else
    Move vehicle to (x,y,z) point
  end if
end if
Goto **START**

2.4 Explosives Tracing and Detection

Algorithm 2 Algorithm for 3-D Mapping of the GPS Traces of Suspicious Metals

Array of (x,y,z) coordinates $\rightarrow A_c$
Array index $\rightarrow I_a$
State variable $\rightarrow S_a$
Program variable $\rightarrow C_p$
$S_a \leftarrow 0$; no change in the present state
$S_a \leftarrow 1$; change in the present state
$C_p \leftarrow 0$; operation is being executed
$C_p \leftarrow 1$; operation is being terminated
Present location point $\rightarrow (x, y, z)$
INITIALIZE:
$X_{\leftarrow} 0$; initial x-axis location
$Y_{\leftarrow} 0$; initial y-axis location
$Z_{\leftarrow} 0$; initial z-axis (or elevated) location
$I_a \leftarrow 0$
START:
Read S_a
if $S_a == 0$ **then**
  $A_c[I_a][0][0] \leftarrow x$
  $A_c[I_a][1][0] \leftarrow y$
  $A_c[I_a][1][1] \leftarrow z$
  $I_a \leftarrow I_a + 1$
end if
if $C_p == 0$ **then**
  for $m < 0; m < I_a; m \leftarrow m + 1$ **do**
    $x \leftarrow A_c[m][0][0]$
    $y \leftarrow A_c[m][1][0]$
    $z \leftarrow A_c[m][1][1]$
    $x \leftarrow x + x_0$
    $y \leftarrow y + y_0$
    $z \leftarrow z + z_0$
    $x_0 \leftarrow x$
    $y_0 \leftarrow y$
    $z_0 \leftarrow z$
    plot (x,y,z) and connect with the previous points
  end for
end if
Goto **START**

The militants can bury improvised explosive devices and landmines in a terrorized area and such malpractice can devastate the entire locality causing huge holistic fatalities of the innocent people and their belongings. To prevent such malicious activities the proposed aerial vehicle based vigilance system is provided with a metal detector circuit module which is attached to a downward inclined portion of the navigator to keep the metal sensing circuitry close to the surface of the landscape. The metal detection technique and implemented sensing system follows the one presented in [2] which contains a submersible monoloop coil and works on the principle of very low frequency electromagnetic induction. The sensing frequency of the detector is about 5.5 kHz which is suitable to detect the emanations from improvised ammos as they contain

very less exposed metallic surface. The sensing depth potential of the configuration is approximately 30 to 35 cm being in elevation from the ground surface. The explorer can detect buried metals from a maximum of 35 cm elevation height measured from the ground surface. The tracing and algorithm for simultaneous 2-D mapping of the detected metal points in the explored region are described in [2]. In this paper, algorithm for consistent 3-D mapping of the scanned surfaces and detected metals has been presented. In the flight mode, aerial vehicle sends (x,y,z) location coordinates of the detected metals where the metal detecting circuit senses the presence of any suspicious material within a maximum depth potential of 35 cm above the ground. Therefore, the z-location values do not exceed 35 cm ground clearance where the x and y-location values are confined within the testing territory. Algorithm 2 shows the implemented approach to map GPS traces of the metals.

3 Results and Analysis

The developed cyber-vigilance system is assessed in different test-beds to ensure the sustainability and reliability of the aerial vehicle based networking signal jammer especially in case of accurate signal blocking activities and stable aerial motion within a confined territory.

3.1 Evaluation of the Signal Blocking Phenomenon

The implemented security system is tested in different environments with different frequency signals with co-existence of the communication signals of 2G, 3G and 4G spectra and wide fidelity (Wi-Fi) and radio frequency (RF) signals. With different combinations of co-existing signals in an arena with a dimension of $10m \times 10m$ with a flight elevation height of $12m$, each time the prototype is tested for 50 times for performance evaluation in terms of accurate blocking (%) and false blocking (%), which are enumerated in Table 1. The evaluated percentages enlisted here are close approximated values.

Table 1: Evaluation of Accurate and False Signal Blocking Phenomena

Test Signals	$B_A(\%)$	$B_F(\%)$
$2G$	96	4
$2G+3G$	86	14
$2G+3G+4G$	85	15
$2G+3G+Wi-Fi(800MHz)$	82	18
$2G+3G+4G+Wi-Fi(800MHz)$	79	21
$2G+3G+Wi-Fi(800MHz)+RF$	78	22
$2G+3G+4G+Wi-Fi(800MHz)+RF$	75	25

Here B_A and B_F are the percent values of accurate and false blocking operations respectively and these are determined as

$$B_A = \frac{N_{BA}}{N_T} \times 100\% \quad (1)$$

$$B_F = \frac{N_{BF}}{N_T} \times 100\% \quad (2)$$

where N_{BA} is the number of accurate blocking operations, N_{BF} is the number of false blocking operations and N_T is the total number of testing operations $[N_T = N_{BA} + N_{BF}]$.
The aerial vehicle is solicited to different elevation heights within a certain territorial boundary of $13.5m \times 13.5m$ and the accurate detection phenomenon is tested for 50 consecutive times for each elevation where each time a composition of $2G + 3G + Wi - Fi(800MHz)$ signals is present there.

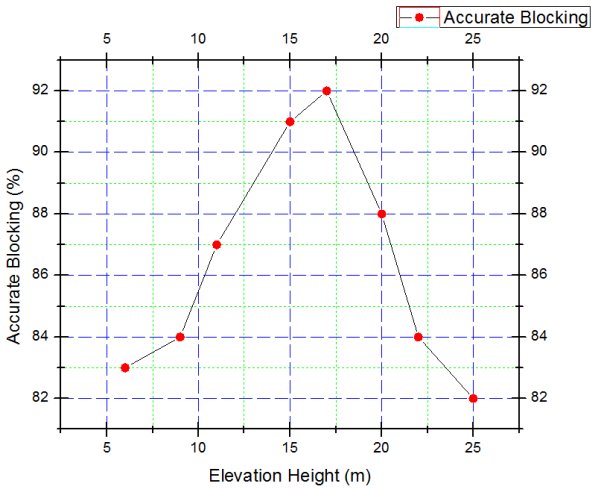


Figure 3: Evaluation of accurate blocking operation in case of different elevation heights

The performance evaluation of accurately blocking signals for different elevation heights is presented in Fig. 3.
Likewise the accurate blocking operation of the signal jammer is tested for different two dimensional territories with a constant elevation height of $12.5m$ for 50 consecutive times. The performance evaluation of accurately blocking signals for different territorial boundaries for a constant elevation height is presented in Fig. 4.

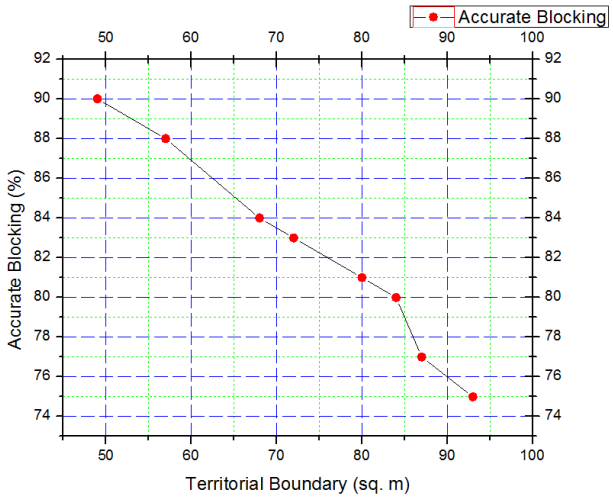


Figure 4: Evaluation of accurate blocking operation in case of different territorial boundaries with a constant elevation height

3.2 Evaluation of the Accurate Metal Tracing Phenomenon

The aerial vehicle is tested for accurate tracing of (x,y) locations of the metals within the testing boundary. In this case, z-locations of the traces are constant for 50 testing attempts. The vehicle in flight mode is controlled to navigate along a 350cm elevation height from the ground surface which is to be scanned to find metal traces. The metal detector is attached to a platform approximately 20cm underneath the main vehicular body and hence the detector senses the metals from a ground clearance of 30cm. Fig. 5 presents the performance evaluation of accurate metal tracing of (x,y) locations within the testing arena.

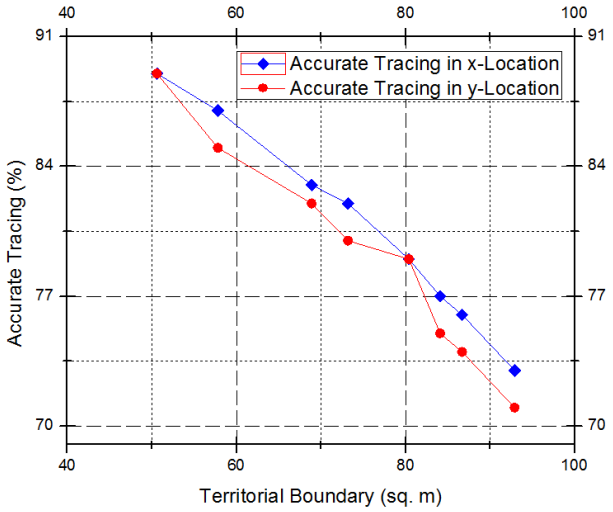


Figure 5: Evaluation of accurate metal tracing for (x,y) locations with a constant z-location value in case of different test arenas

3.3 Evaluation of Deviations in Acceleration

According to [2], the navigation has been controlled in such a manner that the vibration of the aerial structure remains within $-3\ cms^{-2}$ to $+3\ cms^{-2}$ along x and y-axis and $-5\ cms^{-2}$ to $-15\ cms^{-2}$ along z-axis. The linear accelerometer shows some small deviations from the aforementioned standard range when the vehicle is in its flight mode. Fig. 6 presents the (x,y,z) deviations in acceleration of the vehicle. These have been derived for 50 consecutive test runs and the values are approximated from the accelerometer readings.

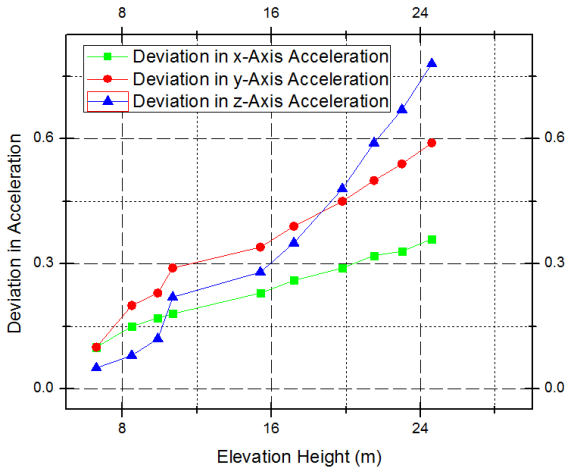


Figure 6: Evaluation of deviations of the vehicle in flight mode from the standard (x,y,z) acceleration values

3.4 Evaluation of Stable Motion within a Confined Territory

The unmanned aerial vehicle is supposed to navigate within a confined region where the signal jammer blocks the networking signals with the provision of video surveillance, map generation and metal detection. The developed system is tested in terms of percentage of accurate confinement to affirm its stable aerial exploration within a certain bounded area. For each different boundary limit, the vehicle is made to explore for 20 times of assessment and the estimated accuracy of confinement in % is enlisted in Table 2.

Table 2: Evaluation of Accuracy of Confinement

Testing Arenas ($m \times m \times m$)	C_A (%)
$7 \times 5.5 \times 5$	93
$8.5 \times 6.3 \times 7.4$	89
$11 \times 7.5 \times 9.8$	87
$14 \times 9 \times 10.2$	82
$16.7 \times 12.5 \times 13$	79

Here

$$C_A = \frac{N_{CN}}{N_{TN}} \times 100\% \quad (3)$$

where C_A is the accuracy of confinement, N_{CN} is the number of confined navigations and N_{TN} is the number of total navigations.

4 Conclusion

This paper articulates a novel aerial vehicle based cyber-vigilance framework capable of networking signal jamming retaliation against militants and terrorists to corner them during an anti-terrorist drive. The unconventional signal jammer consists of a auto tunable band-pass filter module to block communication signals present in the examined area. This surveillance system provides an apprehensive security system which is handy for saving innocent lives stranded in a terrorized region. The aerial prototype is equipped with ammo detection module and has a unique feature of automatic path mapping of the covered region. Accurate signal blocking operation, accurate metal detection, 3-D mapping of the metal traces and stable confined motion with sustainable aerodynamic movement are the prominent features of the proposed work. The test results corroborate the efficacy and authenticity of the developed system for application in defense missions and security drills.

References

[1] D. Chowdhury, M. Sarkar, M. Z. Haider, S. A. Fattah and C. Shahnaz, "Design and Implementation of a Cyber-Vigilance System for Anti-Terrorist Drives Based on an Unmanned Aerial Vehicular Networking Signal Jammer for Specific Territorial Security" in Humanitarian Technology Conference (R10-HTC), Dhaka Bangladesh, 2017. https://doi.org/10.1109/R10-HTC.2017.8288995

[2] S. A. Fattah, M. Z. Haider, D. Chowdhury, M. Sarkar, R. I. Chowdhury, M. S. Islam, R. Karim, A. Rahi, and C. Shahnaz, "An aerial landmine detection system with dynamic path and explosion mode identification features" in IEEE Global Humanitarian Technology Conference (GHTC), Seattle WA USA, 2016. https://doi.org/10.1109/GHTC.2016.7857361

[3] S. A. Fattah, D. Chowdhury, M. Z. Haider, M. Sarkar, M. Refat, G. Rabbi, S. B. Masud, and C. Shahnaz, "Dynamic map generating rescuer offering surveillance robotic system with autonomous path feedback capability" in IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Cebu City Philippines, 2015. https://doi.org/10.1109/R10-HTC.2015.7391872

[4] G. Zhiqiang, L. Zhiheng, H. Zhengxiong, Z. Yonglai, and Z. Xinyuan, "A uhf cmos gm-c bandpass filter and automatic tuning design" in Academic Symposium on Optoelectronics and Microelectronics Technology and 10th Chinese-Russian Symposium on Laser Physics and Laser TechnologyOptoelectronics Technology (ASOT), Harbin China, 2010. https://doi.org/10.1109/RCSLPLT.2010.5615323

[5] P. J. Nicholas, J. C. Tkacheff, and C. M. Kuhns, "Measuring the operational impact of military SATCOM degradation" in Winter Simulation Conference (WSC), Washington DC USA, 2016. https://doi.org/10.1109/WSC.2016.7822342

[6] Y. Zou, J. Zhu, X. Wang, and V. C. M. Leung, "Improving physical-layer security in wireless communications using diversity techniques" IEEE Netw., **29**(1), 42–48, 2015. https://doi.org/10.1109/MNET.2015.7018202

[7] J. Zhu, Y. Zou, and B. Zheng, "Physical-layer security and reliability challenges for industrial wireless sensor networks" IEEE Acce., **5**, 5313–5320, 2017. https://doi.org/10.1109/ACCESS.2017.2691003

[8] M. Nouri, M. Mivehchy, and M. F. Sabahi, "Novel anti-deception jamming method by measuring phase noise of oscillators in lfmcw tracking radar sensor networks" IEEE Acce., **5**, 11455–11467, 2017. https://doi.org/10.1109/ACCESS.2017.2655040

[9] C. Si, H. Sun, M. Sheng, X. Wang, and J. Li, "Physical layer security with hostile jammers and eavesdroppers: Secrecy transmission capacity" in IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia Spain, 2016. https://doi.org/10.1109/PIMRC.2016.7794718

[10] L. Wang, Y. Cai, Y. Zou, W. Yang, and L. Hanzo, "Joint relay and jammer selection improves the physical layer security in the face of csi feedback delays" IEEE Trans. Vehi. Tech., **65**(8), 6259–6274, 2016. https://doi.org/10.1109/TVT.2015.2478029

[11] Q. Yan, H. Zeng, T. Jiang, M. Li, Lou, and Y. T. Hou, "Jamming resilient communication using mimo interference cancellation" IEEE Trans. Info. Fore. and Secu., **11**(7), 1486–1499, 2016. https://doi.org/10.1109/TIFS.2016.2535906

[12] Q. Gao, Y. Huo, L. Ma, X. Xing, X. Cheng, T. Jing, and H. Liu,"Optimal stopping theory based jammer selection for securing cooperative cognitive radio networks" in IEEE Global Communications Conference (GLOBECOM), Washington DC USA, 2016. https://doi.org/10.1109/GLOCOM.2016.7842096

[13] M. A. Mutalib, Z. Zakaria, N. A. Shairi, and W. Y. Sam, "Design of microstrip bandpass filter with electronically tunable notch response" in 26th International Conference Radioelektronika (RADIOELEKTRONIKA), Kosice Slovakia, 2016. https://doi.org/10.1109/RADIOELEK.2016.7477380

[14] L. Hinsz and B. D. Braaten, "A frequency reconfigurable transmitter antenna with autonomous switching capabilities" IEEE Trans. Ante. and Prop., **62**(7), 3809–3813, 2014. https://doi.org/10.1109/TAP.2014.2316298

[15] V. Kumar, R. Mehra, and Shallu, "Reconfigurable band pass filter using kaiser window for satellite communication" in 2nd International Conference on Next Generation Computing Technologies (NGCT), Dehradun India, 2016. https://doi.org/10.1109/NGCT.2016.7877451

[16] S. J. Darak, A. P. Vinod, E. M. K. Lai, J. Palicot, and H. Zhang, "Linear-phase vdf design with unabridged bandwidth control over the nyquist band" IEEE Trans. Circ. and Syst. II: Expr. Brief., **61**(6), 428–432, 2014. https://doi.org/10.1109/TCSII.2014.2319992

[17] C. Taddei, L. Zhuang, M. Hoekman, A. Leinse, R. Oldenbeuving, P. van Dijk, and C. Roeloffzen, "Fully reconfigurable coupled ring resonator-based bandpass filter for microwave signal processing" in International Topical Meeting on Microwave Photonics (MWP) and the 2014 9th Asia-Pacific Microwave Photonics Conference (APMP), Sendai Japan, 2014. https://doi.org/10.1109/MWP.2014.6994485

[18] I. Khoufi, P. Minet, and N. Achir, "Unmanned aerial vehicles path planning for area monitoring" in International Conference on Performance Evaluation and Modeling in Wired and Wireless Networks (PEMWN), Paris France, 2016. https://doi.org/10.1109/PEMWN.2016.7842902

[19] G. Gramajo and P. Shankar, "Path-planning for an unmanned aerial vehicle with energy constraint in a search and coverage mission" in IEEE Green Energy and Systems Conference (IGSEC), Long Beach CA USA, 2016. https://doi.org/10.1109/IGESC.2016.7790077

[20] Y. Liu, R. Lv, X. Guan, and J. Zeng, "Path planning for unmanned aerial vehicle under geo-fencing and minimum safe separation constraints" in 12th World Congress on Intelligent Control and Automation (WCICA), Guilin China, 2016. https://doi.org/10.1109/WCICA.2016.7578482

[21] J. Tao, C. Zhong, L. Gao, and H. Deng, "A study on path planning of unmanned aerial vehicle based on improved genetic algorithm" in 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou China, 2016. https://doi.org/10.1109/IHMSC.2016.182

[22] J. W. Yeol and Y. W. Hwang, "Parametrization of nonlinear trajectory for time optimal 2D path planning for unmanned aerial vehicles" in 2nd International Conference on Control, Automation and Robotics (ICCAR), Hong Kong China, 2016. https://doi.org/10.1109/ICCAR.2016.7486751

[23] F. Lazzari, A. Buffi, P. Nepa, and S. Lazzari, "Numerical Investigation of an UWB Localization Technique for Unmanned Aerial Vehicles in Outdoor Scenarios" IEEE Sens. Journal, **17**(9), 2896–2903, 2017. https://doi.org/10.1109/JSEN.2017.2684817

[24] T. Lee, "Geometric Control of Quadrotor UAVs Transporting a Cable-Suspended Rigid Body" IEEE Trans. Cont. Syst. Tech., **26**(1), 255–264, 2018. https://doi.org/10.1109/TCST.2017.2656060

[25] L. L. Gomes, L. P. Leal, T. R. Oliveira, J. P. V. S. d. Cunha, and T. C. Revoredo, "Unmanned Quadcopter Control Using a Motion Capture System" IEEE Latin America Trans., **14**(8), 3606–3613, 2016. https://doi.org/10.1109/TLA.2016.7786340

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 45-52 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A detailed step-by-step electrical parameters identification method for photovoltaic generators using a combination of two approaches

Selma Tchoketch Kebir*, Mohamed Salah Ait Cheikh, Mourad Haddadi

Laboratoire des Dispositifs de Communication et de Conversion Photovoltaïque, Département d'Électronique, École Nationale Polytechnique, 10 Avenue Hassen Badi El Harrach, Algiers, Algeria

ARTICLE INFO

Article history:
Received: 19 May, 2018
Accepted: 02 July, 2018
Online: 11 July, 2018

Keywords :
Photovoltaic Generators
Data acquisition
Identification
Least Squares
Levenberg-Marquardt
PSO algorithm

ABSTRACT

The object of this paper is to identify the unknown electrical parameters of solar photovoltaic generators in real time, through the application of a novel suggested hybrid method. The identification process is discussed with details, about the four steps of identification. In this issue, the first step describes the experimental data acquisition work done to obtain data from a real photovoltaic system. For the second step, a model of a cell's corresponding electrical circuit is selected. In the third step, the estimation of the parameters 'values is done using two combined optimization approaches, such as Levenberg-Marquardt combined with Particle Swarm Optimization. The fourth step describes the validation of the selected model. The benefit of this work compared to those before, is in the use of real data, in the use of smart optimization technics and the hybridization between two methods, which provides best results.

1. Introduction

This paper is an extension of work originally presented in 2017 at the 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B) [1]. In the order to well detail, explain and develop the previous work presented in this latter conference. In nature, many processes can be modelled by a system of mathematical expressions [2]. Models usually contain some unknown parameters. In the object to find the values of these unknown parameters, some real measurements from the system are required, to be compared with the output of the models proposed through minimizing of some errors [3]. This problem is known as the parameters' identification problem. In these systems, there is the process of modeling and after that, there is an identification process of the unknown parameters for the selected model [4]. In this order, we have used this discipline in a solar photovoltaic (PV) system, in order to identify the electrical parameters of the PV generators. These parameters are very necessary for simulating solar PV production's behavior and therefore control it. As the high non-linearity found in PV cell's models, we present in this work, the combination of two optimization algorithms for identifying the electrical PV generator's unknown parameters from real experimental data. The PV parameters identification process for PV generators is shown in the Figure 1 below.

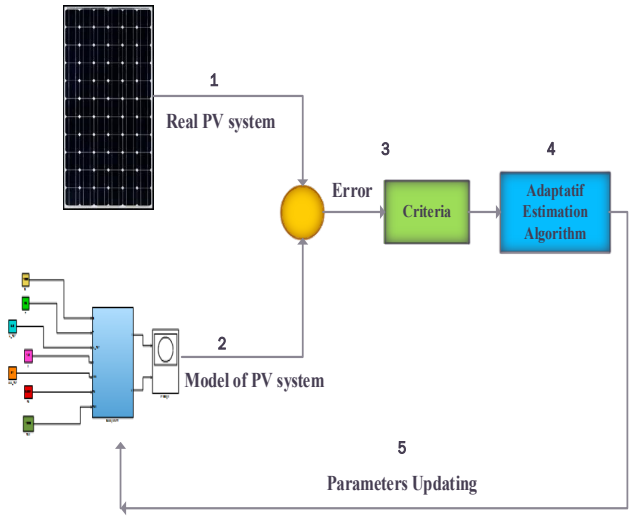


Figure 1. Parameters Identification for PV generators.

The process of PV parameters identification is based on a prediction of the error between the real system and the theoretical model. This error is used with an appropriate algorithm that has to

*Selma TCHOKETCH KEBIR, Ecole Nationale Polytechnique Algiers, Algeria, Email : selma.tchoketch_kebir@g.enp.edu.dz

be modified at each iteration. In this order, the PV parameters are found through the main steps cited below, as illustrated in the above Figure 1:

1- Obtain experimental data from a real PV installed system.

2- Obtain data from a selected model representing a PV system.

3- Minimize the error between the real data and the data of the selected model, using some criteria.

4- Application of an appropriate identification adaptation algorithm.

5- Update the PV parameters values that validate the final selected model representing the real system.

The organization of this paper is as follows: after having introduced the subject in this first section. The second section, gives a state of the art about the PV parameters identification existed methods. Then, the third section focuses on the major steps of PV parameters identification, with more details about each step. Herein, the use of the suggested combined optimization approaches. In the fourth section, a discussion is given concerning the obtained results, before some conclusions in the fifth section.

2. State of the art

In literature, many different methods exist to allow determining the electrical parameters values for the PV generators [5], [6]. Primary, scientific researchers have focuses on the parasitic resistances (series and shunt) determination [7], [8] by the cause of their major influences on the PV performances. Then, it was also observed some influences of all PV electrical parameters on the performances [9], which leads for doing a large number of studies for obtaining their accurate values. The elaborated studies are classified into several categories [5]. Certain methods are classified as theoretical methods such as analytic [10], numeric [11], [12]. Certain other methods are based on an experimental process and are known as identification methods [13], [14], [15]. All methods have their own advantage and inconvenient, to be chosen for using in such application, based on some criteria (complexity, convergence speed, precision …). In this work, we are going to well develop our identification approach.

3. Steps of Photovoltaic Parameters Identification

Parameters identification is an automatic technic, which consists to get a mathematical model, for certain system using some experimental measures. Therefore, the identification process is based on the development of a mathematical representation for a physical system by the use of experimental data [16] [17], through details explained in the main steps cited in the points below of Figure 2.

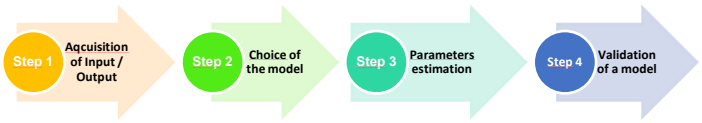


Figure 2. Major steps of parameters identification in systems.

The above major steps of the identification process for PV generators, are well developed in the following points.

2.1. First Step- "Acquiring of Real Data"

A description of our experimental work is giving in this subsection, for obtaining the real PV measured data information [18]. The data acquired issued from a PV system located at our laboratory, (*Laboratoire de Dispositifs de Communication et de Conversion Photovoltaique, LDCCP*). This acquisition work is schematized in Figure 3. The PV system contains a photovoltaic generator, which is composed of four PV panels associated in series, where each panel generate about 55 Watts of power. The measured parameters are the temperature, the solar irradiation, as well as the voltages and currents. In order to collect data, the different components of the proposed PV measurement system are:

- A sensor of temperature 'T', the LM35.
- A sensor of solar irradiation 'G', the pyranometer.
- PV generator associated with the electronic load.

In order to draw the real current-voltage (I-V) curve of PV generators under real conditions, these generators are associated with an electronic load. The process of information reception is made using a computer, which consists of the unit of treatment and control. In this context, we set up an experimental manipulation based on an Arduino-Uno card [19] that allows us the acquisition of data issued from a PV system installed in our laboratory. The transmission of these acquired data to a personal computer is ensured by a USB connection port (serial communication).

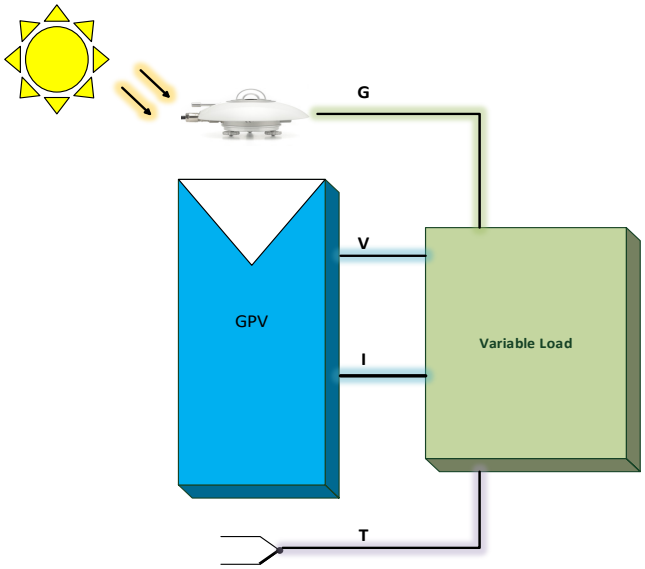


Figure 3. Sensors and PV generator related to a variable load.

The proposed system consists of a set of sensors for measuring both meteorological (e.g. temperature, irradiation) and electrical parameters (photovoltaics voltage and current). The collected data are first conditioned using precision electronic circuits and then interfaced to a PC using a data-acquisition card (Arduino-Uno) [16]. The LABVIEW program is used to further process, display and store the collected data in the PC disk. We have conceived an electronic load based on a MOSFET transistor of power [20]. The variable load allows the measures of a couple of the current-voltage (I–V) along the curve. The obtained current–voltage characteristic (I–V) curve of our photovoltaic generator is

illustrated in Figure 4, which was operated under the real environmental conditions (794.46W/m2, 30.85°C).

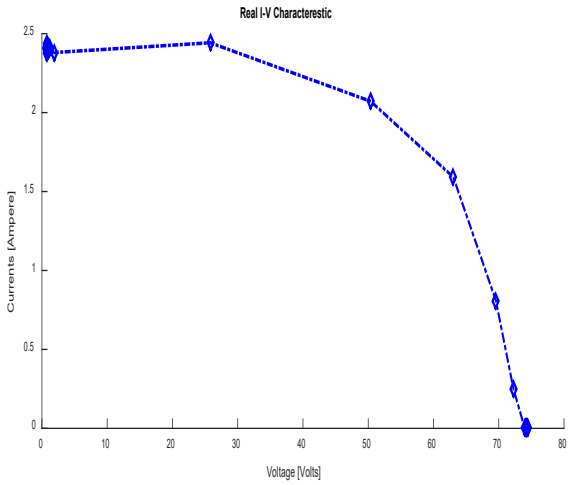


Figure 4. The obtained Current–Voltage experimental curve at (794.46W/m2, 30.85°C) for the four series PV panels of type SM55.

The following *Table I* shows the PV manufacturer's datasheet information of the SM55 PV panel used in this work [21].

Table 1. Specification of Siemens SM55 solar PV panel used, from datasheets under Standard-Test-Conditions : STC (T = 25 °C & G= 1 000 W/m^2).

Parameters		**Mono Crystalline Silicon**
		Panel
Open Circuit Voltage (Volts)	***Voc***	21.7
Short Circuit Current (Ampere)	***Isc***	3.45
Maximum Power Voltage (Volts)	***Vmpp***	17.4
Maximum Power Current (Ampere)	***Impp***	3.15
Maximum Power Point (Watts)	***Pmpp***	55
Number of cells in series	**Ns**	36

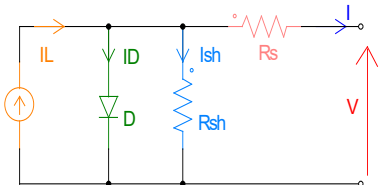


Figure 5. Solar photovoltaic cell's electrical equivalent circuit (The one diode model).

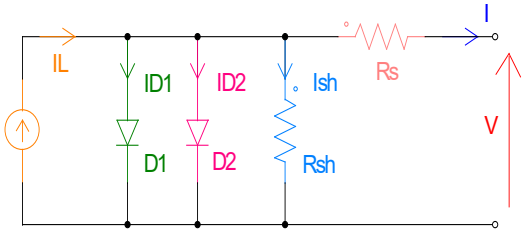


Figure 6. Solar photovoltaic cell's electrical equivalent circuit (The double diode model).

2.2. Second Step- "Selection of the appropriate model"

An accurate mathematical model is an important tool for researchers. The one and double diode PV models, with 5 and 7 unknown parameters respectively, are represented as follow in Figure 5 and Figure 6 [22], [23], [24].

The mathematical equations related the Current-Voltage, (I-V) relationship of the electrical models are given below.

(a) The one diode model

$$I = I_L - I_D - I_{sh} \quad (1)$$

$$I = I_L - I_{ds}.\left(exp\left(\frac{V+R_S.I}{n.V_t}\right) - 1\right) - \frac{V+R_S.I}{R_{sh}} \quad (2)$$

(b) The double diode model

$$I = I_L - I_{D1} - I_{D2} - I_{sh} \quad (3)$$

$$I = I_L - I_{ds1}.\left(exp\left(\frac{V+R_S.I}{n_1.V_t}\right) - 1\right) - I_{ds2}.\left(exp\left(\frac{V+R_S.I}{n_2.V_t}\right) - 1\right) - \frac{V+R_S.I}{R_{sh}} \quad (4)$$

The PV panel contains cells associated in series manners through the mathematical expression below.

$$I = I_L - I_{ds}.\left(exp\left(\frac{V+R_S.I.N_S}{n.V_t.N_S}\right) - 1\right) - \frac{V+R_S.I.N_S}{R_{sh}.N_S} \quad (5)$$

Where:

(a)The one diode model

- I_L: Light current.
- I_{ds}: Diode saturation current.
- n: Diode ideality factor.
- R_s: Series resistance.
- R_{sh}: Shunt resistance.

(b) The double diode model

- I_L: Light current.
- I_{ds1}: Diode saturation current.
- I_{ds2}: Reverse diode saturation current.
- n_1: Diode ideality factor.
- n_2: Second diode ideality factor.
- R_s: Series resistance.
- R_{sh}: Shunt resistance.

The PV Generator contains N_{ms} series branches and N_{mp} parallel branches.

$$I = I_L N_{mp} - I_{ds}.N_{mp}\left(exp\left(\frac{V.N_{mp}+R_S.I.N_S.N_{ms}}{n.V_t.N_S.N_{ms}}\right) - 1\right) - \frac{V.N_{mp}+R_S.I.N_S.N_{ms}}{R_{sh}.N_S.N_{ms}} \quad (6)$$

N_s**:** Number of series cells. N_{ms}: Number of modules in series branches. N_{mp}: Number of modules in parallel branches.

The next Figure 7 shows the PV generator located at our laboratory.



Figure 7. Photovoltaic system mounted at our laboratory.

In our work (Figure 8), the PV generator has N_s of about ***36***, N_{ms} of about ***4*** and N_{mp} is ***1***.

2.3. Third Step- "Parameters estimation"

In this step, the application of the chosen optimization algorithm to identify and obtain the optimal values of PV parameters. The idea is based on a prediction error between the output of the real PV process and the output predicted by the PV model [3], as explained before. This prediction error is used by a Parametric Adaptation Algorithm (A.A.P), which at each iteration will modify the parameters of the model in order to minimize the error, which will be detailed through the next subsections, as schematized in Figure 8.

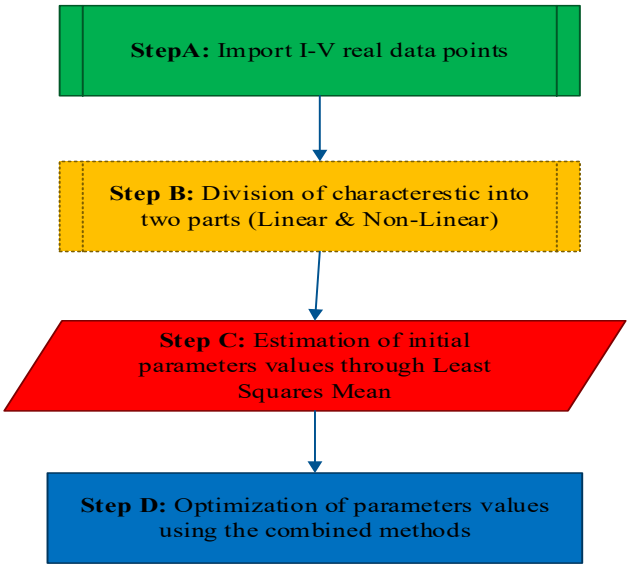


Figure 8. Major steps to estimate, determine and optimize the PV parameters values.

These points will be well explained in the next subsections.

2.3.1. The Least Squares Mean (LSM)

Different approaches allow the **initial estimate** of parameters values such as Least Squares (LS), Least Squares Mean (LSM), Recursive Least Squares (RLS), Recursive Extended Least Squares (RELS) and Newton stochastic (NS) [25]. In an optimization algorithm, the step of initialization of the unknown PV parameters is very necessary for converging to solution. In this regard, we have chosen for obtaining the initial values of our PV parameters the Least Squares Mean (LSM). The benefit of this developed LSM approach is in her non-iterative process. Then, the obtained initial values will be optimized through our developed hybrid approach (Levenberg-Marquardt with PSO). The curve characteristic of the PV generator has two distinct regions as shown in Figure 9, Linear and Non-linear parts. For the linear part, the LSM method is applied merely, while in the non-linear part, it is done through a suitable approximation by a linearization in a logarithmic way [26].

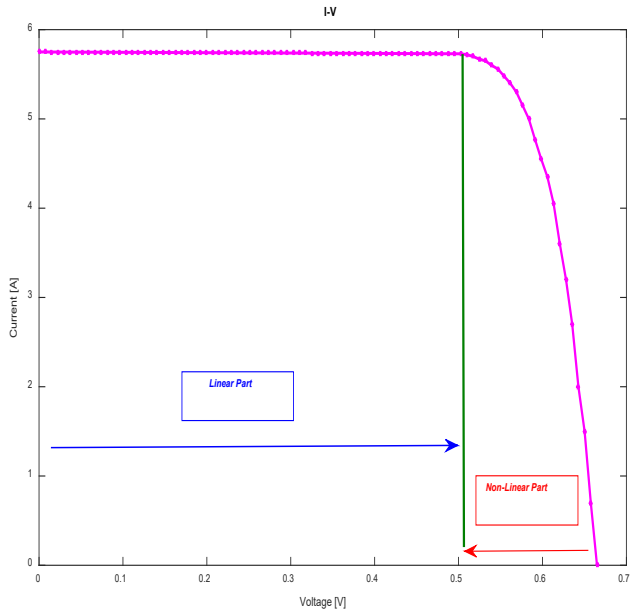


Figure 9. Linear and Non-Linear parts of the I-V characteristic.

The Least Squares Mean approach is based on the minimization of a quadratic function defined as below [14].

$$G = \sum_{i=1}^{n} \varepsilon(i)^2 \quad (7)$$

Where ε is defined as the prediction error between the PV model and the real PV system [14].

$$G = \sum_{i=1}^{n} [I_{Real}(i) - I_{Model}(i)]^2 \quad (8)$$

In order to identify the PV parameters from the real I-V characteristics, we fitted to the best the model to the real (I-V) data, through the minimization of squared errors between the theoretical and experimental curves [27]. In this regard, the cost function used in the optimization process is the Sum of Squared Error (SSE) [28]. Where I_{Model} is the output vector of the PV model, I_{Real} is the PV measured process output and N is the total number of the measured points.

$$G(I,V,X) = \sum_{i=1}^{n} [I_{Real}(i) - f(I,V,X)_i]^2 \quad (9)$$

Where: X is a vector that contains the PV parameters, for example, $\{I_L, I_{ds1}, I_{ds2}, n_1, n_2, R_s, R_{sh}\}$, in the two-diode model. The minimum of Sum Squared Errors "SSE" leads to optimal values of the PV parameters [28]. The minimization of the objective function cannot be done in analytically way due to the strong nonlinearity of the I-V curve characteristics. Therefore, the numerical methods for the nonlinear regression based on the least squares principle are more appropriate to minimize such function.

2.3.2. The Levenberg-Marquardt (LM)

The traditional technic of Levenberg-Marquardt contains inside its process two technics declined from Gradient orders such as the steepest descent and the Gauss-Newton with their complementary features [29]. Generally, the LM approach starts through the steepest descent technic because of her low sensitivity for initial parameters values. Then, after calculating the parameters values and near approaching to the final solution, the Gauss-Newton technic take over and lead to end of the LM process, with a fast convergence. The control parameter δ called damping factor is the responsible parameter in the LM approach that allows switching from steepest descent to Gauss-Newton. The PV parameters to be determined as in [30], are updated at each iteration using the following expression [29].

$$X_{k-1} = X_k - \left[\frac{J'\varepsilon}{J'J+\delta J}\right]_{X=X_k} \quad (10)$$

Where

- J : Jacobian matrix $\left[\frac{\partial f(x)}{\partial x}\right]_{X=X_k}$ which contains a derivative for the function $f(I,V,X)$ regarding each parameter.
- I: Identity matrix.

The damping factor, which his value must be found at each iteration of Levenberg-Marquardt approach, is considered as a crucial factor in the convergence process of the algorithm, for this reason we optimize their value by using PSO approach.

2.3.3. The Particle Swarm Optimization approach (PSO)

This subsection focuses on the evolution study of the function $G(I,\ V,\ X,\ \delta)$, indicated by $G(\delta)$ for X fixed at X_k, regarding many diverse values of the damping factor, at each iteration of the Levenberg-Marquardt algorithm. Herein, it is observed that at each iteration, a presence of various local minimums of $G(\delta)$. For this reason, and in order to achieve the global minimum of $G(\delta)$ corresponding to the best minimizing of the objective function $G(I,\ V,\ X)$, we suggest using the PSO approach. The basic idea of this latter approach was inspired from social behaviour of certain animals such as birds, fishes, and ants and so on. The PSO process is like the natural process existing in groups when they are communicating together, such as a group of birds when they search for migration way or such as ants when they search for food. Herein, it is observed that those animals or insects do not know their best position. Consequently, if a certain member can find the best way to go and achieve the objective, the rest of members, will follow [31]. For mathematical modelling of PSO approach, the population is called a swarm, the individuals are termed as particles and the objective function to be optimized is called fitness. Each particle possesses:

- Its own position

$$P(i+1) = P(i) + V(i+1) \quad (11)$$

- Its own velocity

$$V(i+1) = w * V(i) + c_1 * \left(PL_{best} - P(i)\right) + c_2 * \left(PG_{best} - P(i)\right) \quad (12)$$

- Its neighbourhood (Local and Global)

The local best position is calculated as:

$$PL_{best}(i) = \begin{cases} PL_{best}(i-1) & if\ F(P(i)) \geq F(PL_{best}(i-1)) \\ P(i) & if\ F(P(i)) < F(PL_{best}(i-1)) \end{cases} \quad (13)$$

The global best position is calculated as:

$$PG_{best} = min\left\{F(PG_{best}1), F(PG_{best}2), \dots, F(PG_{best}N)\right\} \quad (14)$$

The following Flowchart, Figure 10, shows the major steps of the PSO's algorithm.

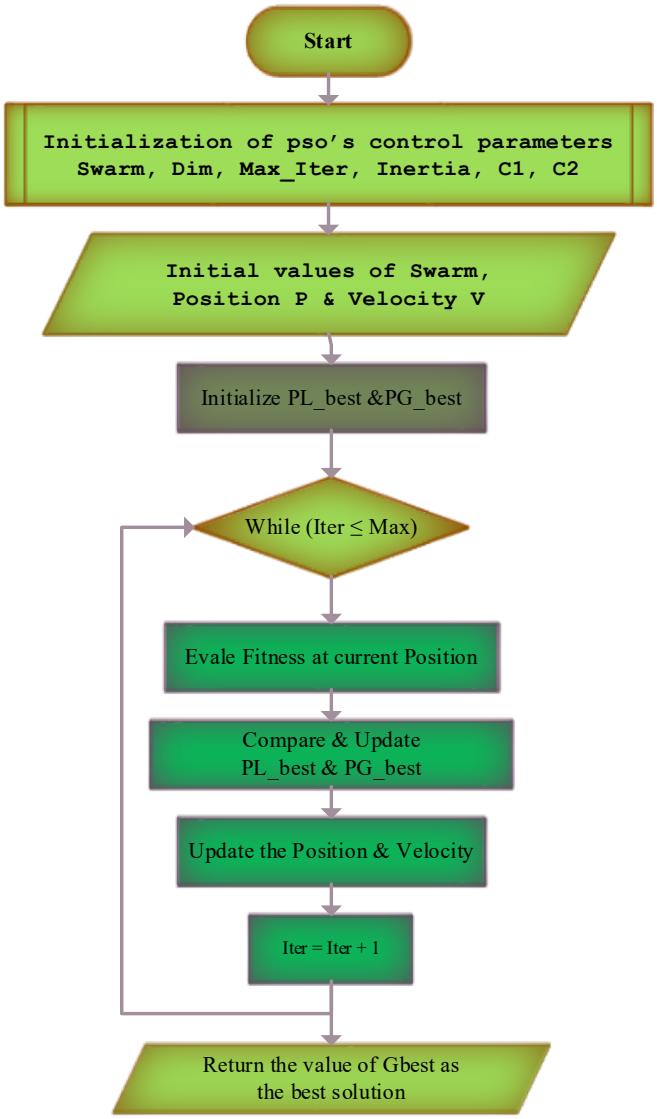


Figure 10. Flowchart of the general steps of PSO algorithm.

This algorithm requires a number of parameters to be set, namely, Inertia W, velocity V, position P, local and global components C_1 & C_2, Number of search agents (the swarm of N visited sites), Maximum number of iteration and the stopping criterion.

2.3.4. *The hybridization of LM with PSO*

The PV identification process, including initialization of PV parameters values through LSM and then optimize their values through the combined Levenberg-Marquardt with Particle-Swarm-Optimization (LM-PSO) developed approach is presented step by step, in what follows Figure 11.

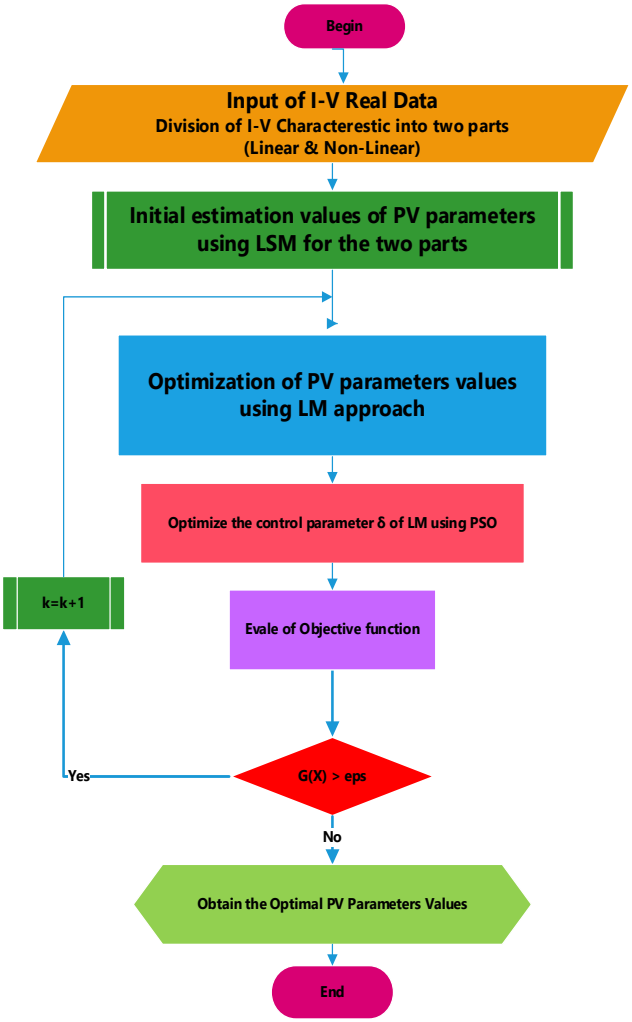


Figure 11. PV parameters identification steps using the hybrid LM approach with PSO approach.

The organigramme of Figure 11 resume the PV parameters identification steps done in this work. After obtaining real data through the acquisition work we used them in a process of algorithms, such as the LSM and the hybrid optimization approach. The advantage of hybridization of two optimization technics is observed in the optimization process of the damping factor of the LM technic, at each iteration, using the PSO technic.

2.4. *Fourth Step "Validation of the model"*

After the estimation process, the obvious query is whether the derived model is adequate for its intended use or not. This is the subjective and overall hard problem of model validation [32]. The sound way to attack the matter is to confront the model with all available type of information, including a priori knowledge, experimental data, and experience of the used model. The selected PV model from the steps before will be validated in this final identification step, after attaining the final obtained results and once fitting all of the entire curve characteristics. The developed approach was tested for other real data of PV cells such as (A-300 of SunPower) as shown in the Figure 12 below, for the one diode model, at STC: Standard-Test-Conditions (T=25°c & G=1000W/m^2).

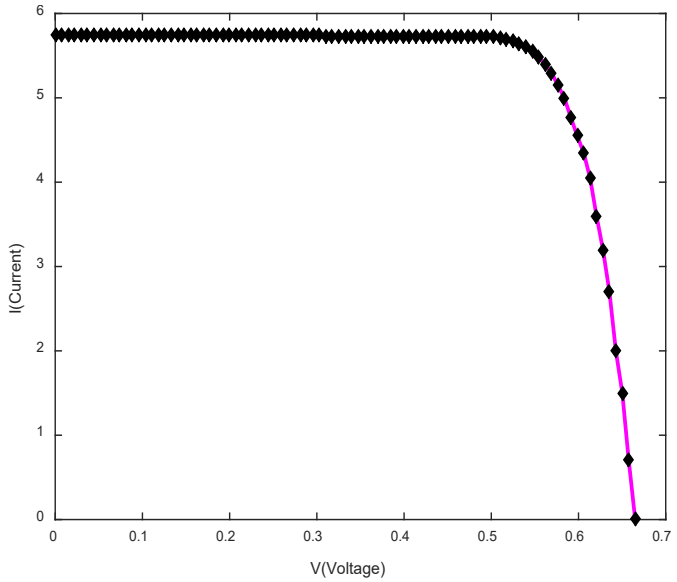


Figure 12. The fitted I-V curve characteristic for the (A-300 SunPower) solar PV cell at STC.

From the above Figure 12, it is clear that once obtaining a good confrontation between the two curves, the model is well chosen to represent the real system.

4. Results & Discussions

This section presents the results obtained for identifying PV parameters values, using the developed combined approach. The following Table II illustrates the results obtained for the electrical PV parameters values of the generator installed in our Laboratory (LDCCP).

Table 2. The obtained parameters values from our developed approach for PV generator.

(a) The one diode model

Parameters	*Obtained values*
I_L	2.45
I_{ds}	3.166611e-07
n	1.279182
R_s	0.036461
R_{sh}	53.271523

(b) The double diode model

Parameters	*Obtained values*
I_L	2.45
I_{ds1}	3.166611e-07
I_{ds2}	3.266611e-07
n1	1.279182
n2	1.379182
R_s	0.036461
R_{sh}	53.271523

The two above tables show that the obtained values for PV parameters models have good values as compared to the values mentioned in literature [33]. The final curves results that validate the model selected for representing our PV generator's behavior, are shown in the fitting I-V curves below, Figure 13.

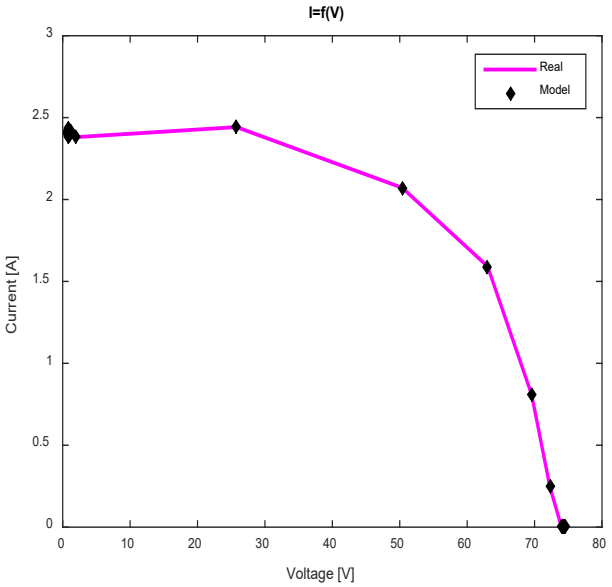


Figure 13. The fitted final I-V curve characteristic for our PV generator (4 associated modules in series of SM55).

It is clear from Figure 13 that the two curves (from the model and the real systems) are well fitted and adjusted. Therefore, the selected model is approximately near of the real behavior of PV generator.

5. Conclusion

In this paper, a detailed study is provided for determining, estimating, identifying and obtaining of the electrical PV parameters values in real time, through combining two algorithms (Levenberg-Marquardt with Particle-Swarm-Optimization). The novelty of this method is that it uses least-squares regressions to initialize the PV parameters values before optimizing their values. The process of identification was elaborated with details, containing all of the four basic steps of identification, as data acquisition, selection of model, PV parameters estimation values and model validation. The obtained results show that the developed approach has the capability to achieve higher PV parameters values with a best precision. The effectiveness is proved by the best approximation obtained between the real data and the data predicted from the curve fitted. This work allows as studying the PV generator's behavior under real conditions, through the detailed steps of the identification process.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The experimental part presented by this paper was done at the LDCCP Laboratory of the Electronic Departement at Ecole Nationale Polytechnique, Algiers, Algeria.

References

[1] S. Tchoketch Kebir, M. S. AIT CHEIKH, M. Haddadi, B. Boumaaraf, "Step by step Parameters Identification for Photovoltaic Generator" in the 5th International Conference on Electrical Engineering – Boumerdes (ICEE-B), Boumerdes, Algeria, 2017.

[2] Karel J. Keesman, System Identification: an introduction, Springer-Verlag London, 2011.

[3] I.D. Landau. System identification in closed loop. Marie Curie Action TOK 3092, 2004.

[4] Gonzalo Cabodevila, Identification des systèmes, École Nationale Supérieure de Mécanique et des Microtechniques, France, 2010.

[5] Ali M. Humada, Mojgan Hojabri, Saad Mekhilef, Hussein M. Hamada, "Solar cell parameters extraction based on single and double-diode models: A review". Renewable and Sustainable Energy Reviews, Volume 56, April 2016, Pages 494-509.

[6] M.A. Hasan S.K. Parida, "An overview of solar photovoltaic panel modeling based on analytical and experimental viewpoint". Renewable and Sustainable Energy Reviews Volume 60, July 2016, Pages 75-83.

[7] J. Thongpron, K. Kirtikara, C. Jivacate. "A method for the determination of dynamic resistance of photovoltaic modules under illumination". Solar Energy Materials & Solar Cells 90 (2006).

[8] Vincenzo d'Alessandro, Pierluigi Guerriero, Santolo Daliento, Matteo Gargiulo. "A straightforward method to extract the shunt resistance of photovoltaic cells from current–voltage characteristics of mounted arrays". Solid-State Electronics 63 (2011) 130–136.

[9] Kenneth, L. Kennerud. "Analysis of Performance Degradation in CdS Solar Cells". IEEE Transactions on aerospace and electronic systems. NOVEMBER 1969.

[10] Valerio Lo Brano, Giuseppina Ciulla, "An efficient analytical approach for obtaining a five parameters model of photovoltaic modules using only reference data". Applied Energy Volume 111, November 2013, Pages 894-903

[11] T.R. Ayodele, A.S.O. Ogunjuyigbe, E.E. Ekoh, "Evaluation of numerical algorithms used in extracting the parameters of a single-diode photovoltaic model". Sustainable Energy Technologies and Assessments Volume 13, February 2016, Pages 51-59.

[12] F.Ghani G.Rosengarten M.Duke J.K.Carson, "The numerical calculation of single-diode solar-cell modelling parameters". Renewable Energy Volume 72, December 2014, Pages 105-112.

[13] Alireza Askarzadeh, Alireza Rezazadeh, "Parameter identification for solar cell models using harmony search-based algorithms". Solar Energy Volume 86, Issue 11, November 2012, Pages 3241-3249.

[14] M.F. AlHajri, K.M. El-Naggar, M.R. AlRashidi, A.K. Al-Othman, "Optimal extraction of solar cell parameters using pattern search". Renewable Energy Volume 44, August 2012, Pages 238-245.

[15] M.Zagrouba, A.Sellami, M.Bouaïcha, M.Ksouri, "Identification of PV solar cells and modules parameters using the genetic algorithms: Application to maximum power extraction". Solar Energy, Volume 84, Issue 5, May 2010, Pages 860-866.

[16] D. Ucinski, Optimal Measurement Methods for Distributed Parameter System Identification: CRC Press; 1 edition (August 27, 2004), 2004.

[17] O. Nelles, Nonlinear System Identification: from classical approaches to neural networks and fuzzy models, 1 ed.: Springer-Verlag Berlin Heidelberg, 2001.

[18] M. Benghanem, "Low cost management for photovoltaic systems in isolated site with new IV characterization model proposed," Energy Conversion and Management, vol. 50, pp. 748-755, 2009.

[19] https://store.arduino.cc/arduino-uno-rev3

[20] H. Belmili, M. S. Ait Cheikh, M. Haddadi, and C. Larbes, "Design and development of a data acquisition system for photovoltaic modules characterization," Renewable Energy, vol. 35, pp. 1484-1492, 2010.

[21] http://www.siemen.co.uk/sm55_sm50.html

[22] Hongmei Tian, Fernando Mancilla-David, Kevin Ellis, Eduard Muljadi, Peter Jenkins, "A cell-to-module-to-array detailed model for photovoltaic panels". Solar Energy, Volume 86, Issue 9, September 2012, Pages 2695-2706.

[23] K. Ishaque, Z. Salam, and H. Taheri, "Simple, fast and accurate two-diode model for photovoltaic modules," Solar Energy Materials and Solar Cells, vol. 95, pp. 586-594, 2011.

[24] Mohsen Taherbaneh ,A. H. Rezaie,H. Ghafoorifard,K. Rahimi Iranian, Iran,M. B. Menhaj &J. M. Milimonfared, "Evaluation of two-diode-model of a solar panel in a wide range of environmental conditions". International Journal of Electronics, Volume 98, Issue 3, 2011.

[25] K. Madsen, H.B. Nielsen, O. Tingleff, Methods for Non-Linear Least Squares Problems, 2nd Edition, April 2004.

[26] K. Bouzidi, M. Chegaar, A. Bouhemadou. "Solar cells parameters evaluation considering the series and shunt resistance". Solar Energy Materials & Solar Cells 91 (2007) 1647–1651.

[27] K.F. Alvina, A.N. Robertson, G.W. Reichc, K.C. Parkd, "Structural system identification: from reality to models". Computers & Structures Volume 81, Issue 12, May 2003, Pages 1149-1176.
[28] Valerio Lo Brano, Giuseppina Ciulla, and Mariavittoria Di Falco. "Artificial Neural Networks to Predict the Power Output of a PV Panel". Volume 2014.
[29] Henri P. Gavin. "The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems". Department of Civil and Environmental Engineering. May 4, 2016.
[30] A. Malaoui, A. Elmansouri, "Deux nouvelles methodes complementaire pour l'extraction optimale des parametres electriques des jonctions". Revue des Energies Renouvelables Vol. 13 N°2 (2010) 199-212.
[31] Satyobroto Talukder. "Mathematical Modelling and Application of Particle Swarm Optimization". Blekinge Institute of Technology, Master of Science, February 2011.
[32] Peter Lindskog. "Algorithms and Tools for System Identification Using Prior Knowledge". Linköping University, September 1994.
[33] Kashif Ishaque, Zainal Salam, Hamed Taheri, Amir Shamsudin. "A critical evaluation of EA computational methods for Photovoltaic cell parameter extraction based on two diode model". Solar Energy 85(9):1768-1779 · April 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 45-51 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Parametric Co-variance Assignment for a Class of Multi-variable Stochastic Uncertain Systems: Output Feedback Stabilization Approach

Qichun Zhang*

School of Engineering and Sustainable Development, De Montfort University, Leicester, LE1 9BH, UK

ARTICLEINFO

Article history:
Received: 08 August, 2018
Accepted: 08 September, 2018
Online: 20 September, 2018

Keywords:
Reduced-order closed-form co-variance model
Parametric co-variance assignment
Output feedback stabilization

ABSTRACT

This paper presents a novel parametric co-variance assignment strategy for multi-variable stochastic uncertain systems. Based upon the explicit parametric design and reduced-order closed-form co-variance model, the variances and co-variances of the system outputs can be assigned artificially using output feedback while the effect of the system uncertainties can be minimized by optimizing the free parameters. In addition, the stability of the closed-loop system has been analyzed and an illustrative numerical example is given to demonstrate the effectiveness of the presented strategy. As a summary, the contributions of this paper include the reduced-order co-variance model, the co-variance error based performance criterion and the parametric control design with stability analysis.

1 Introduction

Co-variance analysis permeates almost all of system theory [1]. Based on the co-variance analysis, the probabilistic properties among random signals can be described. Naturally, the associate co-variance control problem became one of the most significant research topics for multi-variable stochastic systems. Since the system identification, Kalman filtering, stochastic distribution control and fault diagnosis are widely used in practice, the co-variance estimation and control are significant to all of these research areas and applications [2, 3, 4, 5]. In addition, the co-variance is also an ideal tool to analyze the performance of the stochastic systems for the probabilistic decoupling analysis [6, 7] and neural interaction analysis [8, 9].

During the past two decades, the main result of co-variance control is based on the Lyapunov equation while several conditions and controllers have proposed the control design for the co-variance assignment using the determined control signals [10, 11]. However, this controller is designed by Lyapunov equation without closed-form formulation. Since the closed-form model of state co-variance [12] presented in 2007, reduced-order co-variance model [13] has been presented based on eigen-decomposition to solve output co-variance assignment (OCA) problem while the co-variance model was presented for the stochastic system with parametric uncertainty. However, all the existing results did not consider the control design with parametric uncertainties of the stochastic systems. Considering the uncertainties of the parameters, the robust controller has been designed in [14, 15, 16]. All the mentioned controllers can achieve good performance, however these methods have not been used to deal with the co-variance analysis. To the best of our best knowledge, there is no existing solution to the parametric output-feedback co-variance assignment for stochastic uncertain systems with the stability analysis. Therefore, it is significant to develop a simply output co-variance assignment (OCA) control law for the implementation of the complex dynamic multi-variable stochastic system with uncertainties.

In this paper, the stochastic uncertain multi-variable systems have been investigated while the co-variance assignment is very difficult since the uncertainties would affect the stability of the closed-loop systems. Based upon the investigated uncertain multi-variable model, the transformed co-variance model can be obtained firstly, and then the controller and linear observer can be designed and analyzed. In particular, the sufficient conditions are given for the convergence of the observer, the stabilization of the controller and

*Q. Zhang. Email:qichun.zhang@dmu.ac.uk

https://dx.doi.org/10.25046/aj030507

stabilization of the closed-loop system, respectively. In the end, the control input can be obtained by reversing the transformation. Using this control strategy, the design procedure is also given. Furthermore, the parameters of the controller and observer can be optimized using the parametric state feedback [17, 18] and entropy-based performance criterion. Using the presented control strategy, the optimal output feedback control law is obtained for OCA and the performance has been verified by the numerical simulation.

The rest of the paper is organized as follows. In Section 2, the formulation is given including the model formulation, transformed co-variance model and control objective. The parametric control strategy is developed while the convergence of the linear observer, the stabilization of the parametric state feedback controller and the stabilization of closed-loop uncertain system are analyzed in Section 3. Moreover, the parameter optimization and design procedure are also given in this section. Section 4 and Section 5 present the results of numerical simulation and the conclusions, respectively.

2 Formulation

Suppose that the complex industrial dynamic process can be modeled by the following stochastic uncertain multi-variable system.

$$\begin{aligned} dx(t) &= (A+\Delta A(t))x(t)dt + (B+\Delta B(t))u(t)dt + Dd\beta_t \\ y(t) &= Cx(t) \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^s$ and $y \in \mathbb{R}^m$ are the system state vector, input vector and output vector, respectively. β_t is the p-dimensional Wiener process. m, n and s are positive integers while system matrices A, B, C D and parameter uncertainties $\Delta A(t)$, $\Delta B(t)$ are of appropriate dimensions.

Notice that the model (1) can be rewritten as follows:

$$\begin{aligned} dx(t) &= ((A+\Delta A(t))x(t) + (B+\Delta B(t))u(t) + Dw(t))dt \\ y(t) &= Cx(t) \end{aligned} \tag{2}$$

where w is a standard Gaussian white noise. Without loss of generality, assume that the investigated system model (1) satisfies the following assumptions.

Assumption A1: Assume that the Gaussian noise vector $w(t)$ satisfies

$$\begin{aligned} E\{w(t)\} &= 0 \\ E\left\{w(t)w^T(t)\right\} &= 0 \\ E\left\{w(t)w^T(\tau)\right\} &= Q\delta(t-\tau) \end{aligned} \tag{3}$$

where $\delta(\cdot)$ is the Dirac delta function.

Assumption A2: Similar to the assumption of the noise, assume that the control signal is restricted by

$$\begin{aligned} E\{u(t)\} &= 0 \\ E\left\{u(t)u^T(t)\right\} &= 0 \\ E\left\{u(t)u^T(\tau)\right\} &= U(t)\delta(t-\tau) \end{aligned} \tag{4}$$

where $U(t) \triangleq E\left\{u(t)u^T(t)\right\}$.

Based on the definition of the co-variance matrix, the state and output co-variance matrices of the given system are given by

$$\begin{aligned} P_x(t) &\triangleq E\left\{x(t)x^T(t)\right\} \\ P_y(t) &\triangleq E\left\{y(t)y^T(t)\right\} \end{aligned} \tag{5}$$

In addition, the co-variance matrices can be rewritten as follows:

$$\begin{aligned} P_x(t) &= V_x\Lambda_x(t)V_x^T \\ P_y(t) &= V_y\Lambda_y(t)V_y^T \\ U(t) &= V_u\Lambda_u(t)V_u^T \\ Q &= V_q\Lambda_qV_q^T \end{aligned} \tag{6}$$

where Λ_x, Λ_y, Λ_u and Λ_q are real diagonal matrices. V_x, V_y, V_u and V_q are associated orthogonal matrices. All of the matrices are with the same dimensions as the associated vectors.

Therefore, the reduced-order closed-form co-variance model can be obtained following the vectorization operation.

$$\begin{aligned} \dot{\lambda}_x(t) &= (A_{\text{cov}}+\Delta A_{\text{cov}})\lambda_x(t) + (B_{\text{cov}}+\Delta B_{\text{cov}})\lambda_u(t) + D_{\text{cov}}\lambda_q \\ \lambda_y(t) &= C_{\text{cov}}\lambda_x(t) \end{aligned} \tag{7}$$

where λ_x, λ_y, λ_u and λ_q are the diagonal elements of matrices Λ_x, Λ_y, Λ_u and Λ_q, respectively. The coefficient matrices can be calculated using Hadamard product as follows:

$$\begin{aligned} A_{\text{cov}} &= A_\Lambda \circ A_\Lambda \\ B_{\text{cov}} &= B_\Lambda \circ B_\Lambda \\ D_{\text{cov}} &= D_\Lambda \circ D_\Lambda \\ C_{\text{cov}} &= C_\Lambda \circ C_\Lambda \\ \Delta A_{\text{cov}} &= \Delta A_\Lambda \circ \Delta A_\Lambda + 2(A_\Lambda \circ \Delta A_\Lambda) \\ \Delta B_{\text{cov}} &= \Delta B_\Lambda \circ \Delta B_\Lambda + 2(B_\Lambda \circ \Delta A_\Lambda) \end{aligned} \tag{8}$$

while $\Delta A_\Lambda = V_x^T\Delta A V_x$, $\Delta B_\Lambda = V_x^T\Delta B V_u$, $\Delta C_\Lambda = V_y^T\Delta C V_x$, $A_\Lambda = V_x^T A V_q$,$B_\Lambda = V_x^T B V_q$,$C_\Lambda = V_x^T C V_q$ and $D_\Lambda = V_x^T D V_q$. Moreover, reserving the diagonal elements of A_Λ which forms a diagonal matrix A_{diag}, then $A_{\text{cov}} = 2A_{diag}$. Similarly, $\Delta A_{\text{cov}} = 2\Delta A_{diag}$.

Thus, the control objective is to develop a new control strategy to assign the state of the transformed model (7) so that the covariances and variances of the system outputs can be assigned simultaneously.

To achieve the mentioned control objective, the following assumption should be taken into account:

Assumption A3: The pair $(A_{\text{cov}}, B_{\text{cov}})$ is controllable.

Assumption A4: The admissible parameter uncertainties are of the norm-bounded form

$$\begin{bmatrix} \Delta A_{\text{cov}}(t) & \Delta B_{\text{cov}}(t) \end{bmatrix} = M \begin{bmatrix} \Xi_1(t)N_1 & \Xi_2(t)N_2 \end{bmatrix} \tag{9}$$

In Eq. (9), M, N_1 and N_2 denote the structure of the uncertainties which are known real constant matrices

with proper dimensions. $\Xi_1(t)$ and $\Xi_2(t)$ are unknown time-varying matrices which respectively meet the following conditions.

$$\Xi_1^T(t)\Xi_1(t) \le I, \Xi_2^T(t)\Xi_2(t) \le I \tag{10}$$

In addition, the following lemma [19] has been recalled here which can be used to analyze the convergence and stabilization of the presented control strategy.

Lemma 1. *Given any real constant matrices X and Y with proper dimensions. Then there exists a constant* $\xi > 0$, *such that the following inequality holds.*

$$X^T Y + Y^T X \le \xi X^T X + \xi^{-1} Y^T Y \tag{11}$$

3 Control Strategy

The reference co-variance matrix can be rewritten as follows:

$$R = V_r \Lambda_r V_r^T \tag{12}$$

Since the diagonal matrix r can be arranged as vector λ_r, the co-variance assignment problem transfers to state tracking problem using the presented reduced-order co-variance model if we set $V_y = V_r$.

To track the desired state co-variance vector, the integrator should be considered in the control scheme. The error vector $e_y(t) = \lambda_r \lambda_y$ is treated as the extended state and substitutes the error into the closed-loop system.

Then, the closed-loop system in the state-space form can be obtained as follows:

$$\begin{aligned}\begin{bmatrix} \dot{\lambda}_x(t) \\ \dot{\bar{e}}_y(t) \end{bmatrix} &= \left(\bar{A} + \Delta\bar{A}\right)\begin{bmatrix} \lambda_x(t) \\ \bar{e}_y(t) \end{bmatrix} + \left(\bar{B} + \Delta\bar{B}\right)\lambda_u(t) \\ &\quad + \bar{D}\lambda_q + \begin{bmatrix} 0 \\ \lambda_r \end{bmatrix} \\ \lambda_y(t) &= \bar{C}\begin{bmatrix} \lambda_x(t) \\ \bar{e}_y(t) \end{bmatrix}\end{aligned} \tag{13}$$

where

$$\begin{aligned}\bar{A} &= \begin{bmatrix} A_{\text{cov}} & 0 \\ -C_{\text{cov}} & 0 \end{bmatrix}, \bar{B} = \begin{bmatrix} B_{\text{cov}} \\ 0 \end{bmatrix}, \bar{C} = \begin{bmatrix} C_{\text{cov}}^T \\ 0 \end{bmatrix}^T \\ \bar{D} &= \begin{bmatrix} D_{\text{cov}} \\ 0 \end{bmatrix} \Delta\bar{A} = \begin{bmatrix} \Delta A_{\text{cov}} & 0 \\ 0 & 0 \end{bmatrix}, \Delta\bar{B} = \begin{bmatrix} \Delta B_{\text{cov}} \\ 0 \end{bmatrix}\end{aligned} \tag{14}$$

and

$$\dot{\bar{e}}_y(t) = e_y(t) \tag{15}$$

We can further present the closed-loop system in compact format as follows:

$$\begin{aligned}\dot{\bar{x}}(t) &= \left(\bar{A} + \Delta\bar{A}\right)\bar{x}(t) + \left(\bar{B} + \Delta\bar{B}\right)\lambda_u(t) + \bar{D}\lambda_q + \bar{\lambda}_r \\ \lambda_y(t) &= \bar{C}\bar{x}(t)\end{aligned} \tag{16}$$

where

$$\bar{x}(t) = \begin{bmatrix} \lambda_x(t) \\ \bar{e}_y(t) \end{bmatrix}, \bar{\lambda}_r = \begin{bmatrix} 0 \\ \lambda_r \end{bmatrix} \tag{17}$$

Notice that $\bar{D}\lambda_q$ and $\bar{\lambda}_r$ are real constants which means that if the following system is stabilized then the variance and co-variance assignment can be completed.

$$\begin{aligned}\dot{\bar{x}}(t) &= \left(\bar{A} + \Delta\bar{A}\right)\bar{x}(t) + \left(\bar{B} + \Delta\bar{B}\right)\lambda_u(t) \\ \lambda_y(t) &= \bar{C}\bar{x}(t)\end{aligned} \tag{18}$$

Based upon the transformed system model 18, the control strategy can be divided into two parts: observer-based output feedback design and the parametric optimization for uncertainty compensation.

3.1 State feedback design

The linear state feedback controller can be determined by the nominal linear model and the control law is described by

$$\lambda_u(t) = K\bar{x}(t) \tag{19}$$

where the gain matrix K can be obtained by parametric design [17, 13]. In particular, we have

$$\begin{aligned}K &= [W_1 f_1, \ldots, W_m f_m] \\ &\quad \times \left[(\lambda_1^* I - A_1)^{-1} B_1 f_1, \ldots, (\lambda_m^* I - A_m)^{-1} B_m f_m\right]^{-1}\end{aligned} \tag{20}$$

where modified parameter vectors and closed-loop eigenvalues are denoted by $f_1, \ldots, f_m$ and $\lambda_1^*, \ldots, \lambda_m^*$ which can be considered as free parameters. In the case of a common open-loop and closed-loop eigenvalue, the gain matrix K can be determined by the following equations.

$$\begin{aligned}A_j &= \bar{A} + v_j^0 w_j^{0T} \\ W_j &= I - \frac{e_i w_j^{0T} \bar{B}}{w_j^{0T} b_i} \\ B_j &= \bar{B} W_j + v_j^0 e_i^T, j = 1, \ldots, m\end{aligned} \tag{21}$$

where v_j^0 and w_j^0 denote the open-loop eigenvectors and eigenrows of A_j. b_i is the i-th column of $\bar{B}$. e_i is a unit vector while the i-th element is 1. In the other case, no common eigenvalue results in $w_j^{0T} b_i = 0$ which leads to

$$\begin{aligned}A_j &= \bar{A} \\ W_j &= I \\ B_j &= \bar{B}, j = 1, \ldots, m\end{aligned} \tag{22}$$

Substituting control law (19) into the transformed system model (18) yields the closed-loop system:

$$\dot{\bar{x}}(t) = (A_c + \Delta A_c(t))x(t) \tag{23}$$

where $A_c = \bar{A} + \bar{B}K$, $\Delta A_c(t) = \Delta\bar{A}(t) + \Delta\bar{B}(t)K$. Thus the following lemma can be proposed.

Lemma 2. *For the uncertain multi-variable system given by (18), with the assumptions and with the control law given by (19), then there exist two positive constants ε_1 and ε_2, so that the equilibrium $\bar{x}(t)=0$ is stabilized if the following matrix inequality has a positive-definite solution $P=P^T>0$.*

$$A_c^T P + PA_c + \varepsilon_1 N_1^T N_1 + \left(\varepsilon_1^{-1}+\varepsilon_2^{-1}\right)PMM^T P + \varepsilon_2 K^T N_2^T N_2 K + \varepsilon_3 P^T P < 0 \quad (24)$$

Proof. Consider the Lyapunov function candidate as

$$V_c(\bar{x}) = \bar{x}^T(t) P \bar{x}(t), P = P^T > 0 \quad (25)$$

The time derivative of $V_c(\bar{x})$ along the trajectories of (23) is given as follows.

$$\begin{aligned}\dot{V}_c(\bar{x}) &= \bar{x}^T(t) A_c^T P \bar{x}(t) + \bar{x}^T(t) P A_c \bar{x}(t) \\ &+ \bar{x}^T(t)\Delta\bar{A}^T(t) P\bar{x}(t) + x^T(t) P\Delta\bar{A}(t)\bar{x}(t) \\ &+ \bar{x}^T(t) K^T \Delta\bar{B}^T(t) P\bar{x}(t) + \bar{x}^T(t) P\Delta\bar{B}(t) K\bar{x}(t) \quad (26)\end{aligned}$$

Let ε_1 and ε_2 be positive constants, the following matrix inequalities hold using Lemma 1.

$$\begin{aligned}&\bar{x}^T(t)\Delta\bar{A}^T(t) P\bar{x}(t) + x^T(t) P\Delta\bar{A}(t)\bar{x}(t) \\ &= \bar{x}^T(t)(M\Xi_1 N_1)^T P\bar{x}(t) + \bar{x}^T(t) pM\Xi_1 N_1\bar{x}(t) \\ &\le \bar{x}^T(t)(\varepsilon_1 N_1^T N_1 + \varepsilon_1^{-1} PMM^T P)\bar{x}(t) \quad (27)\end{aligned}$$

$$\begin{aligned}&\bar{x}^T(t) K^T \Delta\bar{B}^T(t) P\bar{x}(t) + \bar{x}^T(t) P\Delta\bar{B}(t) K\bar{x}(t) \\ &= \bar{x}^T(t) K^T (M\Xi_2 N_2)^T P\bar{x}(t) + \bar{x}^T(t) PM\Xi_2 N_2 K\bar{x}(t) \\ &\le \bar{x}^T(t)(\varepsilon_2 K^T N_2^T N_2 K + \varepsilon_2^{-1} PMM^T P)\bar{x}(t) \quad (28)\end{aligned}$$

Substituting these inequalities into the derivative of $V_c(\bar{x})$ with Assumption 4, we have

$$\begin{aligned}\dot{V}_c(\bar{x}) &\le \bar{x}^T(t)\left(A_c^T P + PA_c\right)\bar{x}(t) \\ &+ \bar{x}^T(t)\left(\varepsilon_1 N_1^T N_1 + \varepsilon_1^{-1} PMM^T P\right)\bar{x}(t) \\ &+ \bar{x}^T(t)\left(\varepsilon_2 K^T N_2^T N_2 K + \varepsilon_2^{-1} PMM^T P\right)\bar{x}(t) \quad (29)\end{aligned}$$

Since $\dot{V}_c(\bar{x}) < 0$, the proof of lemma 2 is completed. □

3.2 Observer design

Using the linear observer to estimate the states of the model (18), the linear observer can be designed based on the nominal linear model.

$$\dot{\hat{\bar{x}}}(t) = \left(\bar{A} - L\bar{C}\right)\hat{\bar{x}}(t) + L\lambda_y(t) + \bar{B}\lambda_u(t) \quad (30)$$

where the estimated vector can be denoted by $\hat{\bar{x}}$ and L is pre-specified gain matrix of this observer.

Introducing the error of the estimation by

$$e(t) = \bar{x}(t) - \hat{\bar{x}}(t) \quad (31)$$

and substituting the Eq. (30)-(31) to system model (18). The closed-loop model can be described by

$$\dot{e}(t) = A_o e(t) + \Delta A_c(t)\bar{x}(t) \quad (32)$$

where $A_o = \bar{A} - L\bar{C}$. Similar to Lemma 2, Lemma 3 is given as follows.

Lemma 3. *For the uncertain multi-variable system given by (18), with the assumptions and with the linear observer given by (30), then there exists two positive constants ε_1 and ε_2, so that the estimation error $e(t)$ converges to zero if the following matrix inequalities have a positive-definite solution $P=P^T>0$.*

$$A_o P + PA_o < 0 \quad (33)$$

$$\varepsilon_1 N_1^T N_1 + \left(\varepsilon_1^{-1}+\varepsilon_2^{-1}\right)PMM^T P + \varepsilon_2 K^T N_2^T N_2 K < 0 \quad (34)$$

Proof. Consider the Lyapunov function candidate as

$$V_o(e) = e^T(t) P e(t), P = P^T > 0 \quad (35)$$

The time derivative of $V_o(e)$ along the trajectories of (31) is given by the following equation.

$$\begin{aligned}\dot{V}_o(e) &= e^T(t)(A_o P + PA_o)e(t) \\ &+ \bar{x}^T(t)\Delta\bar{A}^T(t) P\bar{x}(t) + \bar{x}^T(t) P\Delta\bar{A}(t)\bar{x}(t) \\ &+ \bar{x}^T(t) K^T \Delta\bar{B}^T(t) P\bar{x}(t) + \bar{x}^T(t) P\Delta\bar{B}(t) K\bar{x}(t) \quad (36)\end{aligned}$$

Similar to the proof of Lemma 2, we have

$$\begin{aligned}\dot{V}_o(e) &\le e^T(t)(A_o P + PA_o)e(t) \\ &+ \bar{x}^T(t)\left(\varepsilon_1 N_1^T N_1 + \varepsilon_1^{-1} PMM^T P\right)\bar{x}(t) \\ &+ \bar{x}^T(t)\left(\varepsilon_2 K^T N_2^T N_2 K + \varepsilon_2^{-1} PMM^T P\right)\bar{x}(t) \quad (37)\end{aligned}$$

which ends the proof □

3.3 Output feedback design

Combining the parametric state feedback controller and the designed observer, the output feedback controller can be obtained for the system (18).

$$\lambda_u(t) = K\hat{\bar{x}}(t) \quad (38)$$

which leads to the closed-loop dynamics as follows.

$$\dot{\bar{x}}(t) = A_c\bar{x}(t) + \Delta A_c(t)\bar{x}(t) - \left(\bar{B} + \Delta\bar{B}(t)\right)Ke(t) \quad (39)$$

Furthermore, the stability of the closed-loop control design can be guaranteed by the following theorem.

Theorem 4. *For the uncertain multi-variable system given by (18), with the assumptions and with the control law given by (38) using the observer (30), then there exists two sets of positive constants $\varepsilon_i, i=1,\ldots,2$ and $\varepsilon_j, j=4,\ldots,7$, so that the equilibrium $\bar{x}(t)=0$ is stabilized if the following matrix inequalities have positive-definite solution $P_1=P_1^T>0, P_2=P_2^T>0$.*

$$\varepsilon_4 K^T B^T BK + \varepsilon_5 K^T N_2^T N_2 K + A_o P_2 + P_2 A_o + \varepsilon_8 P_2^T P_2 < 0 \quad (40)$$

$$\begin{aligned}&A_c^T P_1 + P_1 A_c + (\varepsilon_1^{-1}+\varepsilon_2^{-1}+\varepsilon_4^{-1}+\varepsilon_5^{-1})P_1 MM^T P \\ &+ \left(\varepsilon_6^{-1}+\varepsilon_7^{-1}\right)P_2 MM^T P_2 + (\varepsilon_2+\varepsilon_7)K^T N_2^T N_2 K \\ &+ \varepsilon_3 P_1^T P_1 + (\varepsilon_3^{-1}+\varepsilon_8^{-1})\gamma^2 I_1 + (\varepsilon_1+\varepsilon_6)N_1^T N_1 < 0 \quad (41)\end{aligned}$$

Proof. Consider the Lyapunov function candidate as

$$V(\bar{x}(t), e(t)) = \bar{x}^T(t) P_1 \bar{x}(t) + e^T(t) P_2 e(t) \tag{42}$$

The time derivative of $V(\bar{x}(t), e(t))$ along the trajectories of (39) is shown as follows.

$$\begin{aligned}
\dot{V}(\bar{x}(t), e(t)) &= \bar{x}^T A_c^T P \bar{x} + \bar{x}^T P A_c \bar{x} \\
&+ \bar{x}^T \Delta \bar{A}^T P \bar{x} + \bar{x}^T P \Delta \bar{A} \bar{x} \\
&+ \bar{x}^T K^T \Delta \bar{B}^T P \bar{x} + \bar{x}^T P \Delta \bar{B} K \bar{x} - e^T K^T \bar{B}^T P \bar{x} \\
&- \bar{x}^T P \bar{B} K e - e^T K^T \Delta \bar{B}^T P \bar{x} - \bar{x}^T P \Delta \bar{B} K e \\
&+ e^T(t)(A_o P + P A_o) e(t) \\
&+ \bar{x}^T(t) \Delta \bar{A}^T(t) P \bar{x}(t) \\
&+ \bar{x}^T(t) P \Delta \bar{A}(t) \bar{x}(t) + \bar{x}^T(t) K^T \Delta \bar{B}^T(t) P \bar{x}(t) \\
&+ \bar{x}^T(t) P \Delta \bar{B}(t) K \bar{x}(t)
\end{aligned} \tag{43}$$

Let ε_4 and ε_5 be positive constants, the following matrix inequalities hold using Lemma 1.

$$\begin{aligned}
&- e^T(t) K^T \bar{B}^T P \bar{x}(t) - \bar{x}^T(t) P \bar{B} K e(t) \\
&\le \varepsilon_4 e^T(t) K^T \bar{B}^T \bar{B} K e(t) + \varepsilon_4^{-1} \bar{x}^T(t) P M M^T P \bar{x}(t)
\end{aligned} \tag{44}$$

$$\begin{aligned}
&- e^T(t) K^T \Delta \bar{B}^T(t) P \bar{x}(t) - \bar{x}^T(t) P \Delta \bar{B}(t) K e(t) \\
&\le \varepsilon_5 e^T(t) K^T N_2^T N_2 K e(t) + \varepsilon_5^{-1} \bar{x}^T(t) P M M^T P \bar{x}(t)
\end{aligned} \tag{45}$$

Substituting these inequalities into the derivative of $V(\bar{x}(t), e(t))$ and using Lemma 3, we have

$$\begin{aligned}
\dot{V} \le \bar{x}^T(t) [& A_c^T P_1 + P_1 A_c + (\varepsilon_1 + \varepsilon_6) N_1^T N_1 \\
&+ (\varepsilon_2 + \varepsilon_7) K^T N_2^T N_2 K \\
&+ (\varepsilon_1^{-1} + \varepsilon_2^{-1} + \varepsilon_4^{-1} + \varepsilon_5^{-1}) P_1 M M^T P_1 \\
&+ \left(\varepsilon_6^{-1} + \varepsilon_7^{-1}\right) P_2 M M^T P_2] \bar{x}(t) \\
&+ e^T(t) (\varepsilon_4 K^T \bar{B}^T \bar{B} K + \varepsilon_5 K^T N_2^T N_2 K \\
&+ A_o P_2 + P_2 A_o) e(t)
\end{aligned} \tag{46}$$

which leads to the conditions and the proof has been completed. □

3.4 Parametric optimization

Notice that all the free parameters can be adjusted without changing the stability of the system, which means that the optimization of the parameters can be done for various performance criteria. In this paper, we design the controller based on the nominal model then the optimization can be considered as the uncertainty compensation using the following performance criterion.

$$J = \int_0^t R_1 E\left\{e_y(t)\right\} + R_2 H\left\{e_y(t)\right\} dt \tag{47}$$

where real positive R_1 and R_2 stand for the weights. $H\{\cdot\}$ denotes the entropy. To simplify the performance criterion, the entropy can further be replaced equivalently by information potential [20, 21].

Then the optimal free parameter f_i can be obtained by gradient descent once the eigenvalues λ_i^* are pre-specified.

$$f_{i,j+1} = f_{i,j} + \mu \left.\frac{dJ}{df_i}\right|_{f_i = f_{i,j}} \quad i = 1, \ldots, m \tag{48}$$

where j denotes the optimization searching iteration index. μ stands for the pre-specified step.

Once the control law λ_u is obtained, the control input $u(t)$ for system (1) can be calculated by reversing the transformation. In particular, Λ_U can be obtained by λ_u and we have

$$u(t) = \left[V_u \hat{\Lambda}_u(t) V_u^T\right]^{\frac{1}{2}} \xi(t) \tag{49}$$

where $\xi(t)$ denotes the standard Gaussian white noise.

Remark 1. *The actual control law is non-linear though the transformed system model is linear.*

Remark 2. *Based on the dual principle, the observer gain matrix can be also obtained using proposed optimization approach. Meanwhile, the optimization operation can also be replaced by multi-objective optimization algorithms then the weights can be neglected.*

Remark 3. *Only a few elements of the parameter vectors f_i affect the control performance directly. Therefore, in order to determine the free parameters quickly, trial and error method can be used and the performance criterion can verify the manually selected parameters simply.*

3.5 Design procedure

The procedure of the proposed control strategy is summarized as follows:

Step1 Transfer the system to co-variance assignment model.

Step2 Setup the initial free parameters of the controller.

Step3 Use the numerical approach to optimize the performance criterion (47), by computing the mean, the entropy and gradient descent, then the optimal parameters are obtained.

Step4 Update the feedback gain matrix of the control law (19, 49), and verify it by the conditions of Lemma 2 to guarantee stability of the system, if the conditions hold, then go to next step, otherwise, return to Step 1.

Step5 Obtain the feedback gain matrix of the observer by dual principle and verify it by Lemma 3.

Step6 Verify the optimal parameters by Theorem 4, and if the conditions can be satisfied, then complete the procedure, otherwise, return to Step 2.

Step7 reverse the transformation and obtain the control input by (49).

4 A Numerical Simulation

To verify this new model and the control algorithms proposed in this paper, one numerical example is presented in this section.

The original model can be shown as below:

$$A = \begin{bmatrix} -0.5 & -0.3 \\ 0.1 & -0.2 \end{bmatrix}, B = \begin{bmatrix} 2 & 0.3 \\ 0.1 & 4 \end{bmatrix}$$
$$C = \begin{bmatrix} 0.7 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}, D = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

The co-variance matrix of disturbance white noises is

$$\begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$$

To assign the co-variance matrix, we choose the reference co-variance matrix as

$$\begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}$$

The parametric uncertainties are given as follows:

$$\Xi_1(t) = \Xi_2(t) = I \sin t$$

$$N_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, N_2 = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.1 \end{bmatrix}$$

Simply, we have

$$\lambda_r = \begin{bmatrix} 0.1382 \\ 0.3618 \end{bmatrix}$$

We can further obtain the eigen-vector of the variables as follows:

$$V_r = V_y = V_x = V_u = V_q = \begin{bmatrix} -0.8507 & 0.5257 \\ 0.5257 & 0.8507 \end{bmatrix}$$

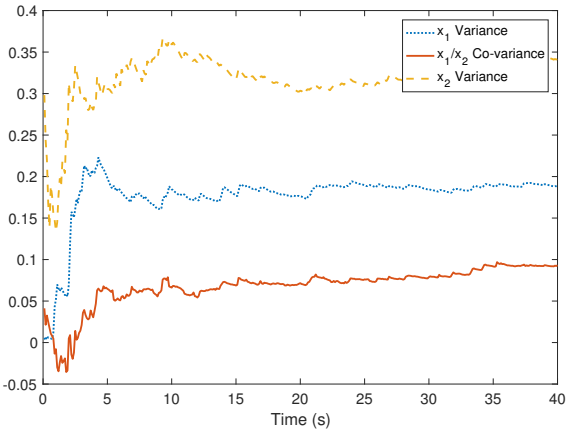


Figure 1. The measured co-variance and variances of the system outputs.

Following the presented control strategy, the simulation results have been shown by Fig. 1-4. Particularly, Fig. 1 indicates the co-variance and variances of the system outputs. Comparing to the reference co-variance matrix, the practical system outputs achieve the assignment with uncertainties. Meanwhile, Fig. 2 shows that the system outputs are stable if the transformed co-variance system design is stabilized while the states and control inputs of the transformed co-variance model have been given by Fig. 3 and Fig. 4. It has been shown that the eigenvalues of the co-variance matrix are tracking the reference eigenvalues of the reference co-variance matrix and the practical control input can be obtained by reversing the transformation using the values of the designed control input of the transformed co-variance model.

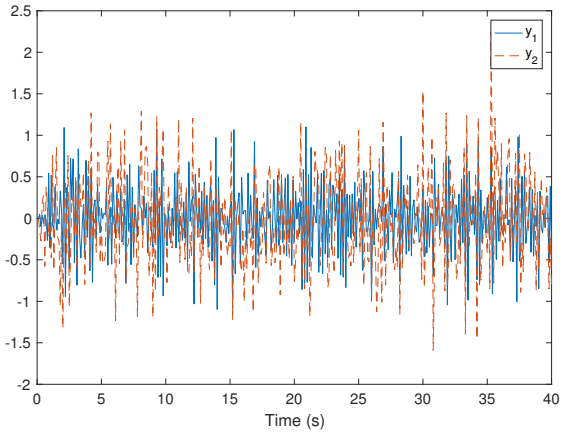


Figure 2. The outputs of the system.

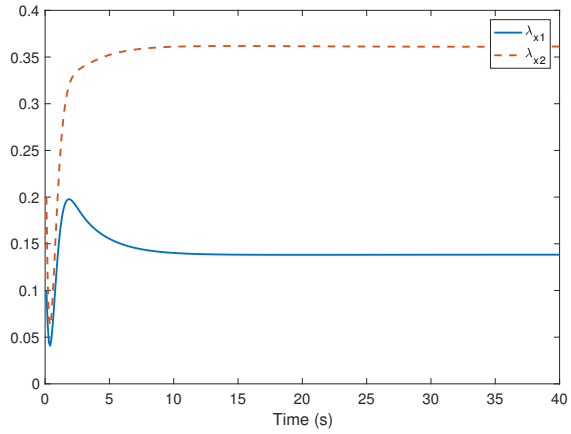


Figure 3. The state assignment of the transformed co-variance model.

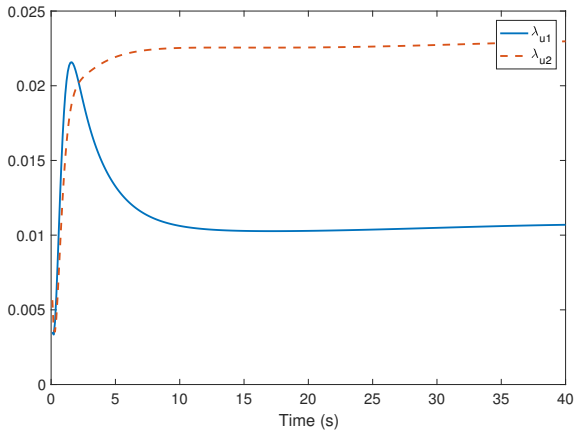


Figure 4. The designed control input of the transformed co-variance model.

5 Conclusion

This paper investigates the co-variance assignment strategy for a class of multi-variable stochastic uncertain systems. Combining the reduced-order covariance model and parametric feedback, the control strategy is obtained by output feedback stabilization. In particular, the transformed model is given firstly with the extended co-variance assignment error. Then the linear observer is designed to estimated the extended state of the transformed model. After that, the output feedback is obtained via parametric optimization. Meanwhile the theoretical analysis is given to guarantee the robustness, stabilization and convergence of the closed-loop systems. Based on the results of the numerical simulation, the effectiveness of the presented control strategy has been verified while the control objectives have been achieved. Since the covariance assignment is widely used in practical systems, such as paper-making process, the industrial applications using the presented control strategy will be the potential extension as a future work.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgement The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] A. Hotz and R. E. Skelton, "Covariance control theory," *International Journal of Control*, vol. 46, no. 1, pp. 13–32, 1987.

[2] K. Liu, K. Li, and C. Zhang, "Constrained generalized predictive control of battery charging process based on a coupled thermoelectric model," *Journal of Power Sources*, vol. 347, pp. 145–158, 2017.

[3] M. Ren, J. Zhang, M. Jiang, M. Yu, and J. Xu, "Minimum (*hphi*)–entropy control for non-gaussian stochastic networked control systems and its application to a networked dc motor control system," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 1, pp. 406–411, 2015.

[4] L. Guo, L. Yin, H. Wang, and T. Chai, "Entropy optimization filtering for fault isolation of nonlinear non-gaussian stochastic systems," *IEEE Transactions on Automatic Control*, vol. 54, no. 4, pp. 804–810, 2009.

[5] L. Yao, J. Qin, H. Wang, and B. Jiang, "Design of new fault diagnosis and fault tolerant control scheme for non-gaussian singular stochastic distribution systems," *Automatica*, vol. 48, no. 9, pp. 2305–2313, 2012.

[6] Q. Zhang, J. Zhou, H. Wang, and T. Chai, "Minimized coupling in probability sense for a class of multivariate dynamic stochastic control systems," in *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*. IEEE, 2015, pp. 1846–1851.

[7] ——, "Output feedback stabilization for a class of multivariable bilinear stochastic systems with stochastic coupling attenuation," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2936–2942, 2017.

[8] Q. Zhang and F. Sepulveda, "A statistical description of pairwise interaction between nerve fibres?" in *Neural Engineering (NER), 2017 8th International IEEE/EMBS Conference on*. IEEE, 2017, pp. 194–198.

[9] ——, "A model study of the neural interaction via mutual coupling factor identification," in *Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE*. IEEE, 2017, pp. 3329–3332.

[10] K. Yasuda and R. E. Skelton, "Covariance controllers: A new parameterization of the class of all stabilizing controllers," in *American Control Conference, 1990*. IEEE, 1990, pp. 824–829.

[11] K. M. Grigoriadis and R. E. Skelton, "Minimum-energy covariance controllers," *Automatica*, vol. 33, no. 4, pp. 569–578, 1997.

[12] H. Khaloozadeh and S. Baromand, "State covariance assignment problem," *IET control theory & applications*, vol. 4, no. 3, pp. 391–402, 2010.

[13] Q. Zhang, Z. Wang, and H. Wang, "Parametric covariance assignment using a reduced-order closed-form covariance model," *Systems Science & Control Engineering*, vol. 4, no. 1, pp. 78–86, 2016.

[14] Z. Wang, B. Huang, and H. Unbehauen, "Robust reliable control for a class of uncertain nonlinear state-delayed systems," *Automatica*, vol. 35, no. 5, pp. 955–963, 1999.

[15] T. Shen and K. Tamura, "Robust h/sub/spl infin//control of uncertain nonlinear system via state feedback," *IEEE Transactions on Automatic Control*, vol. 40, no. 4, pp. 766–768, 1995.

[16] Q. Zhang and X. Yin, "Observer-based parametric decoupling controller design for a class of multi-variable non-linear uncertain systems," *Systems Science & Control Engineering*, vol. 6, no. 1, pp. 258–267, 2018.

[17] G. Roppenecker, "On parametric state feedback design," *International Journal of Control*, vol. 43, no. 3, pp. 793–804, 1986.

[18] G. Roppenecker and J. O'reilly, "Parametric output feedback controller design," *Automatica*, vol. 25, no. 2, pp. 259–265, 1989.

[19] I. R. Petersen, "A stabilization algorithm for a class of uncertain linear systems," *Systems & Control Letters*, vol. 8, no. 4, pp. 351–357, 1987.

[20] Q. Zhang and A. Wang, "Decoupling control in statistical sense: minimised mutual information algorithm," *International Journal of Advanced Mechatronic Systems*, vol. 7, no. 2, pp. 61–70, 2016.

[21] Y. Zhou, Q. Zhang, H. Wang, P. Zhou, and T. Chai, "Ekf-based enhanced performance controller design for nonlinear stochastic systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 1155–1162, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 46-55 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Holistic Access Control and Privacy Infrastructure in Distributed Environment

Uche Magnus Mbanaso[1,*], Gloria A Chukwudebe[2]

[1]*Centre for Cyberspace Studies, Nasarawa State University, Keffi, Nigeria*

[2]*Department of Electrical & Electronic Eng., Federal University of Technology Owerri, Nigeria*

ARTICLE INFO

Article history:
Received: 18 August, 2018
Accepted: 27 October, 2018
Online: 01 November, 2018

Keywords:
Internet
Internet of Things
Cyber- physical system
Digital trust,
Confidentiality,
Privacy
Access Control
Distributed Environment

ABSTRACT

This article discusses IoT security in situations whereby devices do not share the same security domains, which raises security, privacy and safety concerns. It then presents an Access Control and Privacy infrastructure for addressing these concerns in the context of distributed environments. IoT deployments allow billions of connected physical devices to collect, process and share data; collaborate and cooperate in automating tasks in an unrivaled fashion. However, security and safety are still top major fears that demand holistic approach, particularly when devices do not share the same digital trust. This is not a surprise, as a revolutionary system, IoT comes with inherent vulnerabilities, threats and risks like most other computing and data processing systems. Conversely, when security breaches or compromises occur, it is most likely to have a far-reaching and upsetting consequences that extends traditional concerns. The fact that IoT can be deployed in plethora of application scenarios; means that end-to-end security should be treated contextually and in a dynamic manner. Consequently, these concerns; trust, confidentiality, and privacy at the IoT application stack need to be addressed robustly. Thus, in this article, a novel distributed access control infrastructure based on configurable policy constructs is presented. The infrastructure provides a mechanism for gradual negotiated release of provable attributes to dynamically build trust before protected resources are made available. In this configuration, IoT transaction parties can express their Capabilities (competences, features, etc.) and Requirements (rules and provable attributes required to access the capabilities) as the basis for sharing data or collaboration in solving business problems.

1. Introduction

This paper is an extension of work originally presented during the 13th International Conference on Electronics, Computer and Computation (ICECCO) in 2017 [1]. The deployment of Internet of Things (IoT) is developing in many areas and contexts. Its deployment spans across diverse spaces and is anticipated to continue to extend beyond present expectations [2]. In some deployments, the range of IoT devices that may work together or share data are unlikely to belong to a single (or the same autonomous) security domain [3]. By security domain, we basically mean a collection of connected entities or applications that are part of a specific digital trust infrastructure (or administered by common cryptographic policy), i.e. Public Key Infrastructure (PKI) security arrangement for authentication, authorization and session management. Invariably, when devices, which are members of different security domains want to collaborate in providing business solutions, it raises trust, confidentiality and privacy challenges in a variety of application contexts. This demands fresh security requirements as threats posed by cyber, physical and human factors span beyond traditional risk landscapes. However, trust, confidentiality and privacy have received substantial attention in the literature in different contexts [4][5][6][3][7].

Notwithstanding, IoT has different set of physical and virtual (or logical) fresh crop of security issues that varies, and are contextually multifaceted. That is, IoT security, privacy and safety may be seen from a variety of surfaces including those specific to the device, cloud computing, mobile apps, network interfaces, software, physical and access control. While a number of these security areas, can be addressed specifically by vendors and/or manufacturers, yet some have to be addressed in application context and real-time, and cannot be on the basis of 'one solution fits all'. Arguably, it requires security infrastructure

*Uche Mbanaso, Centre for Cyberspace Studies, Nasarawa State University, Keffi, Email: ozara.oru@gmail.com

that is adaptable and flexible, taking into consideration the business or solution environments. It therefore aptly suggests that IoT operational environments can be highly contestable with several attack opportunities for intruders. This raises strong digital trustworthiness as germane for treating identity and access management in emerging smart environments. The assurance of how data is collected, processed and shared, entails an obligation to mutually respect contractual data access agreements that met security principles and privacy.

A typical example can buttress our point. A smart vehicle arriving a city would like to request certain city-based data from available connected city IoT systems [2]. But a secure conversation will demand that these city devices cannot wittingly disclose data without first ascertaining the trustworthiness of such third party entity. It is assumed here that the smart vehicle belongs to another security domain and has no existing digital trust affiliation with the city's security domain[8]. For security and safety purpose, both parties should exchange information based on the ability to trust each other[9][10]. It implies that access and data security as well as safety of these connected entities must be reciprocally assured. Inversely, the smart vehicle may equally not be ready to disclose its profile to the city systems straight away, and the city systems cannot assume that the smart vehicle's mission is harmless, and thus share data with the vehicle or conduct operations together. In either direction, both parties have security, privacy and safety issues of concerns. In this regard, critical challenges can be inferred as follows:

- Unauthorised activities of hostile entities to compromise security and safety of IoT;
- Activities of friendly parties to disregard mutually contractual agreements to violate security and safety of devices, resources or underpinning infrastructure.

To this extent, IoT security and safety landscapes are still evolving issues constrained by encumbrances that have both socio-economic and security impetus. Furthermore, privacy and trust are subjective to cultural perceptions with unpredictable degree of individual capacity and expectations [11]. Unlike trust that is by no way regulated, privacy rights are subjective to some sort of regulations, especially by legislation, security principles, procedures, ethics, etc. in a variety of countries [12][13]. Aside, it is expected that a blanket access cannot be allowed to typical IoT systems, particularly for safety reasons; besides the need to secure sensitive resources and/or attributes. This, uniquely makes a further strong case for a symmetric security infrastructure at application stack that supports fine-grained policy in decision-making and authorization.

Therefore, it is obvious that IoTs are likely to operate in a variety of security domains and without pre-established digital trust but may have to share data, and where possible work together to address common business issues but in a secure and trusted manner. Thus, it is incumbent that identity and access management at application stack are critical requirement for treating security, privacy and safety. In this light, we present a bilateral symmetric and configurable policy-based infrastructure to address this critical application layer security in IoT distributed systems. This infrastructure uses Obligation of Trust (OoT) protocols[3] that allows reciprocated interexchange of policy constructs described as *Requirements* and *Capabilities* to gradually establish dynamic trust before making available protected information or performing some mutual tasks. Traditional solutions assume a form of digital trust based on simple use of username/password pair, which is highly susceptible to a variety of threats [14][15][16]. This is not ideal in many of IoT solution space, particularly in distributed environments.

Our solution is novel in many respects. First, it offers a real-time mutual treatment of security and privacy using configurable policy constructs that permits both parties in transaction to reciprocally take access decision based on their individual security requirements and capabilities. Second, it is a departure from one-way protection perception whereby only the security concerns of the party providing services is considered. Third, it is a highly scalable access control mechanism that has the capability to deal with present and future threats through robust and extensible rule constraints. Fourth, by addressing trustworthiness in privacy protection in a unified fashion, the infrastructure provides mechanisms for accountability, trust-based digital evidence as basis for dispute resolution, which is a critical requirement for IoT security and safety.

The rest of the paper is organized as follows: Section II reviews related works while Section III presents threat analysis and challenges in the context of privacy, trust and confidentiality. Section IV describes the novel Distributed Access Control Infrastructure while Section V presents discussions on the novel infrastructure. Section VI concludes the paper.

2. Review of Related Works

IoT requires a holistic approach to solve security, privacy and safety concerns in a particular security layer. This may entail combining technical, procedural and legal controls to minimize the severity of risks associated with access and availability of protected data as well as intellectual or proprietary property [1][3]. IoT operations take place at application stack where systems collect, store, analyze and share data. In some cases, sensitive attributes of service requesting parties are required to perform authorization. Undisputedly, privacy is most often considered from simple legal statements without automation of enforceable technical measures. Meanwhile IoT application level security is similar to those faced by other computing application space, existing identity and access control models can be adapted to suit IoT environments. However, the complexity is that an autonomous security domain may have hundreds or even thousands of connected objects with sensors and actuators to manage. This is the differentiator, which makes IoT risk landscape differ significantly. To this extent, the challenge before us is how to extend and adapt existing models and controls to address numerous IoT security issues. In the section that follows, security models and standards that influenced our solution are reviewed.

2.1. IoT Reference Model

The increasingly broad adoption of IoT devices span wide area of use cases across multiple business domains including smart cities, smart manufacturing, smart agro, smart parks, smart hospitals, smart patient supported living solutions, etc.[17].

In smart cities, for instance, IoT sensors can focus on sensing the environment on some crowded areas. For instance, sensors can be used to ascertain air quality among others in an effort to monitor particular densely inhabited city zones periodically [18]. However,

sensor data can be spoofed or can become attack vector to facilitate a particular threat. In this context, understanding the tenets of security, privacy and safety issues is incumbent to the different IoT security layers. Thus, the IoT reference models create a common understanding of operational layers, features and functionality of IoT, which can help in insightful conceptualisation security architecture [18]. More importantly, it is instructive to note that no single security layer is a complete solution. However, there are plethora of IoT reference architectures, which helped to conceptualise IoT security[18][14][19].

2.2. Federated Identity Management (FIM)

Simply, the Federated Identity Management (FIM) is an infrastructure model used to associate identity information and attributes of entities across trusted several security domains [20]. The approach provides a mechanism for "single sign-on" in a fashion that allows transaction parties to obtain trusted access tokens from their local Identity Provider in order to be allowed access to outside services in a confederated manner [10][21]. An example of the FIM is the OpenID [22]. Usually, FIM is a classical transient trust built by using username/password pair to authenticate to a party's local Identity Provider (IdP) while this IdP issues and communicates signed access control assertions or tokens to the service providing party.

In some deployments, Attribute Authority (AA) is an integral part of FIM developed to provide a much more resilient access control engine as opposed to simple authentication provisions [23][24]. Typically, AA is simply, a trusted repository for secure storage of attributes/properties of parties commonly used in Attribute-Based Access Control (ABAC) infrastructure [20]. In some use cases, FIM attempts to distinguish authentication operations from authorization process on the basis of separation of security duties.

Although FIM offers user convenience and efficiency in managing identity provisions, users and relying parties, the use of username/password pair makes it defenseless against numerous threats. More so, issuance of access tokens is not on itself sufficient to guarantee the behaviour of a transaction entity. However, managing vast IoT identities has been raised as also a contending issue due to the anticipated volume of IoTs. Thus, FIM is potentially well suited for managing IoT identities, and as an integral part of a distributed access control infrastructure.

2.3. eXtensible Access Control Markup Language (XACML)

The XACML is a standard access control policy construct developed by the Organization for the Advancement of Structured Information Standards (OASIS), that provides collective framework for specifying a range of access control rules [25]. It has its foundation from eXtensible Markup Language (XML), and presents an extensive access control structure and encoding schemes to describe fine-grained access control rules as well as message level request-response construct that allow constituent part to work together in distributed access control operations. It exemplifies a modular infrastructure that is loosely coupled based on functionality and application domain, in a manner that allows them to be hosted independently.

2.4. Security Assertion Markup Language (SAML)

SAML is a very powerful and extensible language based on XML scheme specifically developed for the exchange of access control information from one transaction party to another. Usually, an identity provider (a SAML issuer or SAML authority) makes one or more assertion statements about a principal or entity in an opaque string, which is communicated to a consuming party, typically a service provider to grant access to the subject described on the digitally signed assertion [26][27]. The relying party decides to trust the SAML issuer based on some pre-existing trust relationship provided by digital certificate, which asserts that the subject is trustworthy.

From purely technical perspective, SAML assertion is the primary standard used by most single sign-on (SSO) schemes, even the FIM. The XML structure has an Issuer element that describes the SAML authority; the *Signature* section that holds the *signature block,* which encapsulate the PKI information of the issuer, algorithms and transforms as well as the resulting digital signature, etc. The *Subject* element encapsulates the identity of the subject; the *Condition* element describes obligatory conditions as an additional rule constraints. The *Assertion Statement* specifies the assertion context including authentication, attribute, authorization decision, or other user-defined constructs that can facilitate access control.

2.5. Obligation of Trust (OoT) Protocol

The Obligation of Trust Protocol (OoT) provides an innovative symmetric access control protocol as described in [7][28], which illustrates a bilateral and symmetric method that combines digital trust negotiation and access control operations for the treatment of security and privacy protections based on enforcement of mutual policy rules between two or more parties in distributed application environments. Ideally, the OoT protocol allows two or more transaction parties to interexchange policy constructs contained in *Requirements* and *Capabilities* in real-time. The OoT SAML request message described as a *Notification of Obligation* (NoB), first notifies the services requesting party the conditions for accessing its resources expressed as *Requirements* and its available services or features in *Capabilities.* The response message after execution of *Matching Algorithms* is the assurances that describes the fulfillment of each other's conditions contained in the *Requirements* policy element. The response message is characteristically the *Signed Acceptance of Obligations* (SAO). The details of OoT access control protocol that demonstrates how parties in conversation can use SAML Obligation of Trust Assertion can be found in [3][7].

3. Threat Analysis and Challenges

Fundamentally, to understand operational IoT security and safety issues, all layers of IoT must be considered and thoroughly assessed. Thus, the five security goals i.e. confidentiality, integrity, availability, authenticity, and non-repudiation should form the basis to assess threats. Consequently, in assessing these threats, three classic IoT system threats are described as follows:

- **A Target of an Attack**

Conventionally, an IoT device is potentially exposed to many threats faced by a typical computing system, particularly at network and application layers. It implies that IoT can suffer data breach or the device can be degraded, which can result to violation of confidentiality (or privacy) and integrity, as well as denial of service (availability). Most IoT systems have in-built application servers that equally face the same security challenges as traditional web servers[29]. OWASP[30], described ten top categories of IoT vulnerabilities that can be exploited by a hostile

party. Thus, threats can materialize through evading authentication provisions due to weak configurations and associations to the extent that it is too difficult to repudiate (non-repudiation) the nefarious actions. The extent to which this can happen depends on the mission, capability and the motivation of a hostile party.

- **A Tool for an Attack**

The composition of an IoT device includes sensors and actuators, implying the potential to be manipulated intelligently to distribute nefarious programmes or become an integral part of a malicious network that can take part in Distributed Denial of Service (DDoS) attack to cause unavailability. Likewise, operationally, an IoT can simplify a variety of attacks as sensors and actuators can conveniently become attack vectors. A malicious party can leverage unpretentious IoT operations for illegitimate purposes. In this context, such attacks may include undercover use of IoT engine to perpetuate cybercrime, financial fraud or cyberwarfare.

- **Incidental to an Attack**

This type of threat becomes possible when an IoT ecosystem is indirectly involved in an attack (i.e. stealthily supports criminal activities such as when in itself it is used to store data for criminal activities). It infers that possibly, an IoT can expedite an attack to occur much quicker by leveraging its power or functionality, or operational processes, which can make an attack more challenging to detect and attribute thereby causing non-repudiation attack.

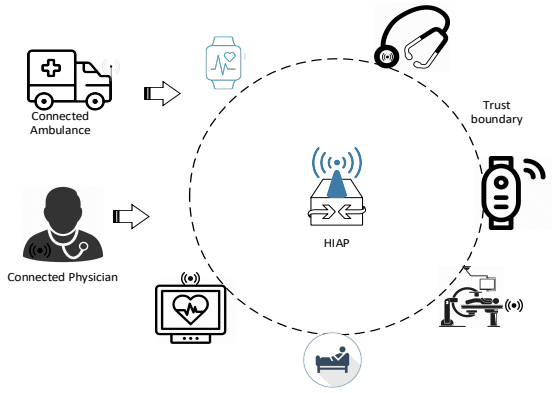


Figure 1: Typical IoT Use Case.

Already, cybercriminals have leveraged the inherent vulnerabilities in IoT engines to cause major disruptions, especially the Distributed Denial of Service (DDoS) attacks, which exploited Domain Name System (DNS) requests [31]. With the estimated 20 billion connected physical objects by 2020, and the explosion of industrial internet of things, recruiting thousands of connected devices to cause DDoS may be trivial. The foregoing typically suggests that a trusted party authorized to gain access to an IoT engine, can misuse it by employing the device to carry out other functions than originally programmed. For instance, GDPR[32] prescribes that personal data be used only for the initially stated purposes [13]. This provision makes privacy a contractual responsibility that must be respected by transacting IoT entities. Similarly, trust relationship is the requirement that attempts to guarantee the expected behaviour i.e. the hope that an IoT entity will behave reciprocally and responsibly without impairment to the other party. Thus, in practice, this mutuality may be too difficult to achieve.

Potentially, IoT poses different set of threats and risks in diverse environments and contexts, which should be addressed dynamically and in perspective [4]. For instance, in healthcare scenario, a Physician may have the need to work in several hospitals, of which her digital trust is not provided by the same or single security domain. It implies that this kind of use case requires that a high level of trust be established, privacy to be guaranteed, and confidentiality to be kept as well as the assurance of accountability and non-repudiation [33]. In healthcare environment, diagnosis, monitoring and assessment of patients may require significant number of devices interconnected by Heterogeneous IoT Access Point (HIAP) to collaborate and cooperate to solve patient's problems. In the same vein, the Physician's IoTs without previous trust relationship, may be required to interact with other IoTs within the environment. In this scenario, for convenience sake, the Physician IoTs should discover and connect automatically to the same HIAP or gateway without recourse to manual configuration. Another simple real-world use case is a connected ambulance that brings a patient to a smart hospital environment, under the characteristics of the mission (or emergency), the ambulance should be able to discover and automatically connect to relevant devices to accomplish its mission without manual configurations.

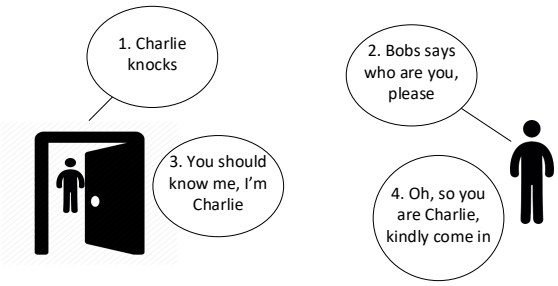


Figure 2: Trust Dialogue between Bob and Charlie.

Figure 1 depicts a typical smart healthcare environment. In many IoT systems, the manufacturers can provide a sort of security abstraction, which other security features can be derived, such features cannot by default solve application layer security that is usually contextual.

Like other computing devices, inbuilt security at abstraction layer cannot address application layer security and privacy out of the box, especially in distributed environments. In practical sense, security, privacy, and trust are not static security requirements. The implication is that these issues must be treated in context and instantaneously too. Moreover, IoT systems are probably going to expose services through Application Programmable Interface (API), this further reinforces the requirement for dynamic security, privacy and safety solutions that should be configurable[2]. To further provide insight to underlying concepts, we examine trust, privacy and confidentiality individually, and in perspective.

3.1. Trust Context

Building digital trust in typical IoT in distributed environment raises fresh security issues. In digital space, building trustworthiness is vital, and can then be built between physical

objects, physical objects and people, physical objects and systems as well as systems. Thus, trust is a critical factor in distributed IoT environments that must be examined holistically. Theoretically, a simple trust dialogue between Charlie and Bob can be used to demonstrate generally, the subtlety of trust as a concept, shown in Figure 2:

Step 1: Charlie arrives at Bob's door and knocks;
Step 2: Bob says 'who are you, please?'
Step 3: Charlie answered, 'you should know me, I'm Charlie';
Step 4: Bob says "oh, so you are Charlie, kindly come in. In this case Bob allowed Charlie because he seems to recognize the voice of Charlie and anticipated to see him.

Examining this simple trust dialogue, there is a potential that Bob can open the door and see an imposter (who imitated Charlie's voice) instead of Charlie. This simplistic example, can be the basis to further discuss three important variables associated with trust namely: behaviour, reputation and expectations. Bob has merely trusted the statement based on known reputation and behaviour of Charlie with the anticipation that he will remain trustworthy. This modest example suggests there is inherent risk factors in the general concept of trust, thereby buttressing the point that current trust models provided by Public Key Infrastructure (PKI) are sufficient to guarantee trust in IoT environments. It further underscores the fact that providing security and safety features in IoT distributed systems require a sort of arrangement that gives the communicating parties to gradually establish more trust based on other attributes beyond PKI provisions.

In literature [34][23][35], digital trust is well researched, and provides the mechanism to verify and validate trust relationships, privileges, claims, identity attributes and information, etc.; giving the identity consuming party the opportunity whether to rely on the real-time assertions of proving party or not, based on the extended properties defined in the rule constraints.

3.2. Direct vs Indirect Trust

Traditionally, digital trust is simply based on either direct or indirect (or transitive) trust relationships. In a typical IoT system, access to resources can be granted based on verification and validation of pre-existing trust relationships that authenticates a party requesting a service. Generally, this can be referred to as a direct trust, a form of shared secret, such as username/password pair or digital certificates, etc.; which is usually created offline between parties prior to communications as depicted in Figure 3(a). Figure 3(b) illustrates the concept of indirect trust whereby a service provider requires to verify and validate the assertions made by a party requesting service but there is no existing digital trust relationship between them. Thus, for a secure conversation, a trusted intermediary must prevail to vouch for the requesting party in a manner that a relying party can trust its assertions. Such examples as practiced today include signing into other third party online applications using Facebook or Twitter accounts.

In highly sensitive and safety critical IoT applications, simple trust models as described above is flawed substantially, which can easily be fooled by malicious parties. Based on the analysis already presented, different application contexts in distributed IoT environments, will require full-proof digital trust built gradually by reciprocated negotiation of verifiable attributes to assure security, privacy and safety.

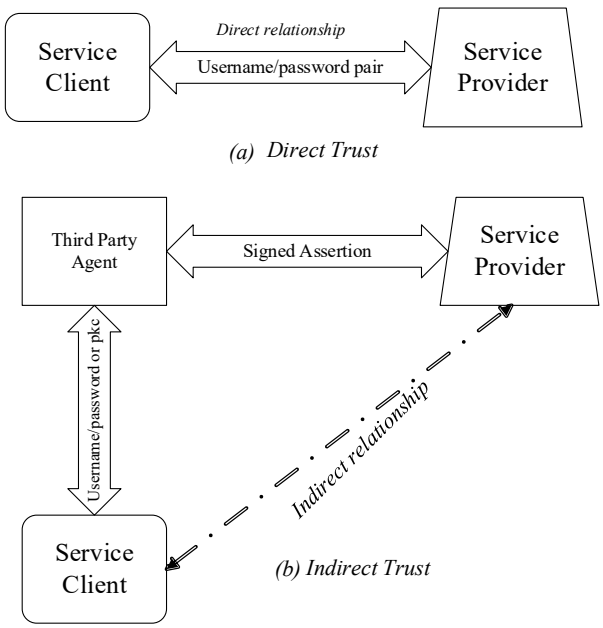


Figure 3: Classical Trust Models

3.3. Privacy and Confidentiality Context

In many instances, privacy and confidentiality still remain a misunderstood concepts. It is not surprising as the terms are closely related. Conversely, in distributed application environments, where two or more actors are involved, privacy and confidentiality operationally are difficult to guarantee. For example, during access control phase, data of privacy value may be shared between these actors i.e. from requesting party's domain to the relying party, yet this privacy data may be disclosed by an intermediary party of another security domain. Data is fluid, and once shared with a third party, exercising full control thereafter becomes uncertain. While confidentiality can obviously be addressed by access restriction and/or encryption, privacy is subjective to trust expectations. So, it can be inferred that confidentiality is a means to ensure privacy protection especially when data is at rest but once the same data is passed unto a third party, then privacy may be eroded.

It is certain that personally identifiable information shared during transactions can be stored by either parties, to which the collector purportedly proclaims controls on behalf of the data owner. This raises privacy concerns, the data subject may lose track of the parties holding its data, and has no option but to rely on the facts of promise statements that the information will be given adequate privacy safeguards and protection. However, the new European Union General Data Protection Regulation (GDPR) has altered data privacy protection[13][36]. In the same wise, it may seem obviously that GDPR legal rules may have put stringent proscriptions but monitoring compliance and conformance is still operationally, a challenging task. Consequently, it suggests that in real-time, managing identity and access management, requires interacting parties to ensure that vouching for trustworthiness is cryptographically signed.

Above, entails that trust, privacy and confidentiality are strongly related and require a homogenous infrastructure to address them concurrently. To this extent, important questions can be raised to stimulate design assessment as follows:

(i) How can transaction entities account for their actions when privacy attributes are compromised or breached?
(ii) What are the technical mechanisms that can monitor how privacy data are accessed, shared and processed?
(iii) What is the assurance that a party can keep privacy promises made to another party, support and safeguard proportionately by suitable operational means?
(iv) Is there a technical mechanism to guarantee that transfer and processing of privacy information conforms to relevant standards and regulations, and its subsequent processing by second level third party?
(iv) What are the mechanisms to handle conflicts and risks? Is there a valid channel to handle and resolve conflicts that supports strong digital evidence?
(v) How can the liable parties be determined in multifaceted data breach involving several actors?
(vi) Is there any difficult-to-repudiate digital evidence that is admissible in courts of law to support assertions in an event of disagreement?

Imperatively, these questions can form open issues that challenges the research community and the need to find optimal solutions to address complex security, privacy and safety posture of IoT threat and risk landscapes, especially in the wake of increasing value of data [36][32]. Furthermore, it is an acknowledged fact that technology alone cannot answer all of the questions hypothesized above. As a consequence, it is remarkable to state that suitable governance, regulation and compliance, conflict resolution and assurance mechanisms, are vital inputs, which buttress the point that there is a strong interplay between technology, policy and law in solving privacy equations.

Notwithstanding, robust technical infrastructure has significant role in responding to the above named issues. Technically speaking therefore, dealing with real-time security, privacy and safety of connected devices operationally, require a flexible and distributed infrastructure that supports configurable policy constructs to manage IoT risks based on informed and preferred decisions.

4. Distributed Access Control Infrastructure

To design applicable access control infrastructure for IoT in distributed systems requires thorough examination of the various actors in a typical IoT conversations. As illustrated in Figure 4, there are likely to be multiple actors from different security domains that can interoperate in classical IoT service deployments. This is assumed on the ground that one security provider may be unsuitable for identity and access control to authenticate and validate security assertions that can be trusted across some high profile IoT distributed environments [12].

For instance, access to IoT systems in classical medical environment may require personal attribute of Medical Consultants from the Medical Council as well as a referrer attributes from a City Council as the basis to share or disclose resources. Equally, in a smart city, a rule requirement may entail that for vehicles to interact with city cameras for example, the vehicle license plate number as well as insurance certificate may need to be authenticated before access is allowed to protected resources. In this typical case, the attribute providers may unlikely be part of a single security domain. It implies that in distributed application scenarios, IoT access requires a robust and scalable infrastructure to treat security, privacy, and safety dynamically and in trusted fashion.

Figure 4 clearly illustrates conceptual view of access control entities in distributed environments. It shows entities and the various responsibilities as well as data flows. It is widely acknowledged that IoT is resource constrained for now; buttressing the fact that process consuming access control operations such as evaluation of access control policies may not be executed within IoT system. As such, light weight access control operations e.g. such as enforcement of decisions can be carried out in IoT systems, while other functions be delegated to trusted external parties.

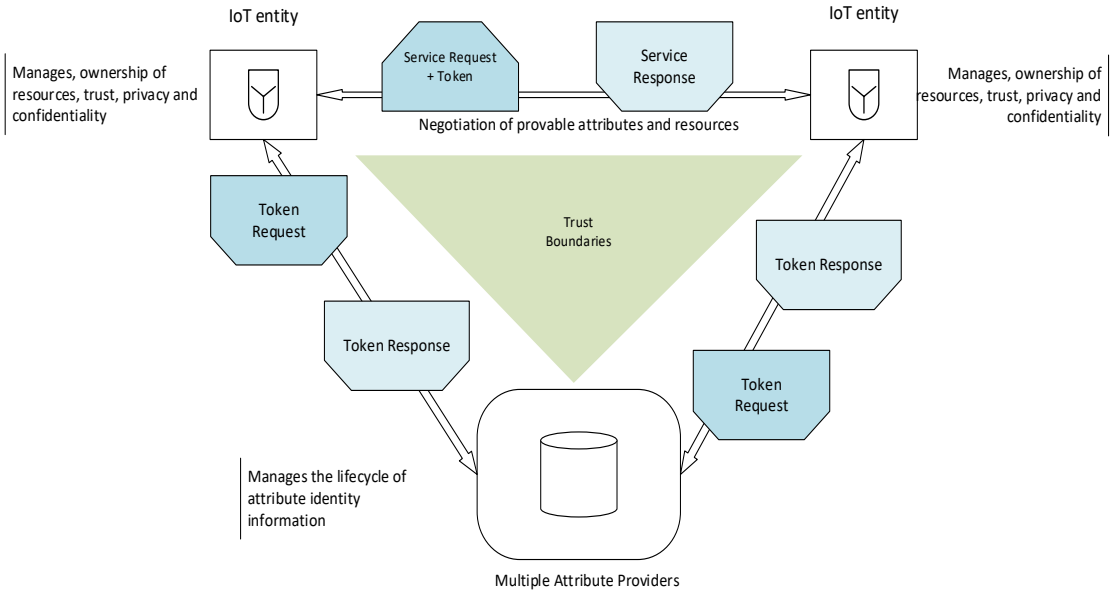


Figure 4: Conceptual View of Access Control Entities in a Distributed Environment.

4.1. Distributed Access Control Infrastructure for IoT

Figure 5 depicts Distributed Access Control Infrastructure that integrates components of FIM, Identity/Attribute Authority (IAA), and Obligation of Trust (OoT). The infrastructure components are loosely coupled in a distributed manner to allow interoperability and flexibility in deployment due to resource constrained IoT environment. The infrastructure consists of three logical subunits grouped according to areas or separation of concerns. The gatekeeper is tightly coupled to IoT application stack, which comprises Context Handler (CH) and Policy Enforcement Point (PEP) components of XACML. These components can programmatically be part of IoT system through its web service interface. The CH formulates or interprets specific application context data in a required format during conversations. Similarly, the PEP engine is responsible for the enforcement of access control decisions arriving at the gatekeeper after interpretation by CH. The Policy Decision Point (PDP) and Policy Information Point (PIP), still component of XACML provides decision point where serious policy *Matching Algorithms* are implemented. The Identity/Attribute Provider (I/AAP), which is derived from the concept of FIM supplies trusted attributes of entities to facilitate

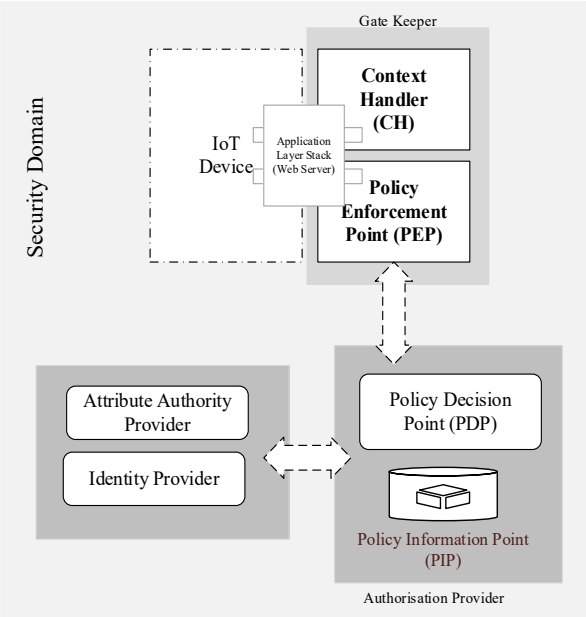


Figure 5: Distributed Access Control Infrastructure.

decision making by PDP. These subunits can then be hosted anywhere in the cloud to provide identity and access control functions. This infrastructural arrangement clearly shows the importance of IoT device belonging to a security domain where there is prevailing trust relationship with the authorization service for the solution to be feasible.

4.2. SMAL OoT Protocol

Figure 6 illustrates a protocol sketch between IoT entities in distributed systems, a way to mutually interexchange SAML OoT Assertion messages in order to decide whether resources can be shared either way [3].

The sequence of interactions are explained in the following steps:

1) A classical smart vehicle (SV) arriving a city sends a service request to Smart City Systems (SCS).
2) The request is intercepted by SCS Security gate keeper (CH/PEP), which constructs and sends SAML OoT containing Notification of Obligation (NoB) context.
3) The SV gate keeper intercepts, constructs and responds with its SAML OoT that contains its NoB.
4) The SCS CH constructs another SAML OoT that contains both NoBs and sends to its PDP for processing and decision.
5) The SCS's authorization engine based on the policy attributes sends SMAL OoT message to SV's IdP/AA requesting verification of identity/attributes of SV.
6) The SV's IdP/AA sends a corresponding SAML OoT *Response* message containing signed identity/attributes requested or makes a fresh request to the sending party (5 & 6 can iterate number of times depending on the trust negotiation configuration).
7) The SCS's authorization engine using the policy sets (NoBs) and based on 6 response, performs the *Matching Algorithm* to determine access decision.
8) The SCS sends SAML OoT *Response Message* that contains Signed Acceptance of Obligations (SAO) based on 7.
9) The SV's sends corresponding SAML OoT message that contains its SAO to SCS.
10) Then, SCS sends the requested resources to SV.

Note: Usually, direct trust can be the basis to initiate the negotiation, which usually, is a form of simple authentication; but not illustrated in the diagram.

4.3. Obligation of Trust Policy Architecture

Technically speaking, IoT in distributed setting is operationally complex and sophisticated, especially in interconnected and integrated application environments where applications talk to applications. To mutually treat trust, confidentiality and privacy, requires a configurable policy set that is robust and scalable. In this context, a policy construct that provides *Requirements* element and *Capabilities* element that permit each party to expressively and granularly describe its obligations and expectations is presented. Conversely, for IoT transaction parties to collaborate together in real-time, they can mutually express and interexchange the policy constructs containing *Requirements* and *Capabilities* in order to treat trust, confidentiality and privacy concurrently as illustrated in Figure 7. As demonstrated, party A's *Requirements* must match with party B's *Capabilities*, and similarly, party B's *Requirements* must match party A's *Capabilities* in a typical access request evaluation. This mutual evaluation gives each party, using granular expressive rules, the preference to decide and balance the sharing of resources in comparison to their mutual benefits. This construct when combined with *Digital Signature* solves confidentiality, integrity, authenticity and non-repudiation, thereby meeting basic security goals in a typical secure transaction.

In summary:

i. *Requirements* element is used to express a party's obligations (or commitments), it would expect another party requesting for a resource to fulfil before such a resource can be made available;

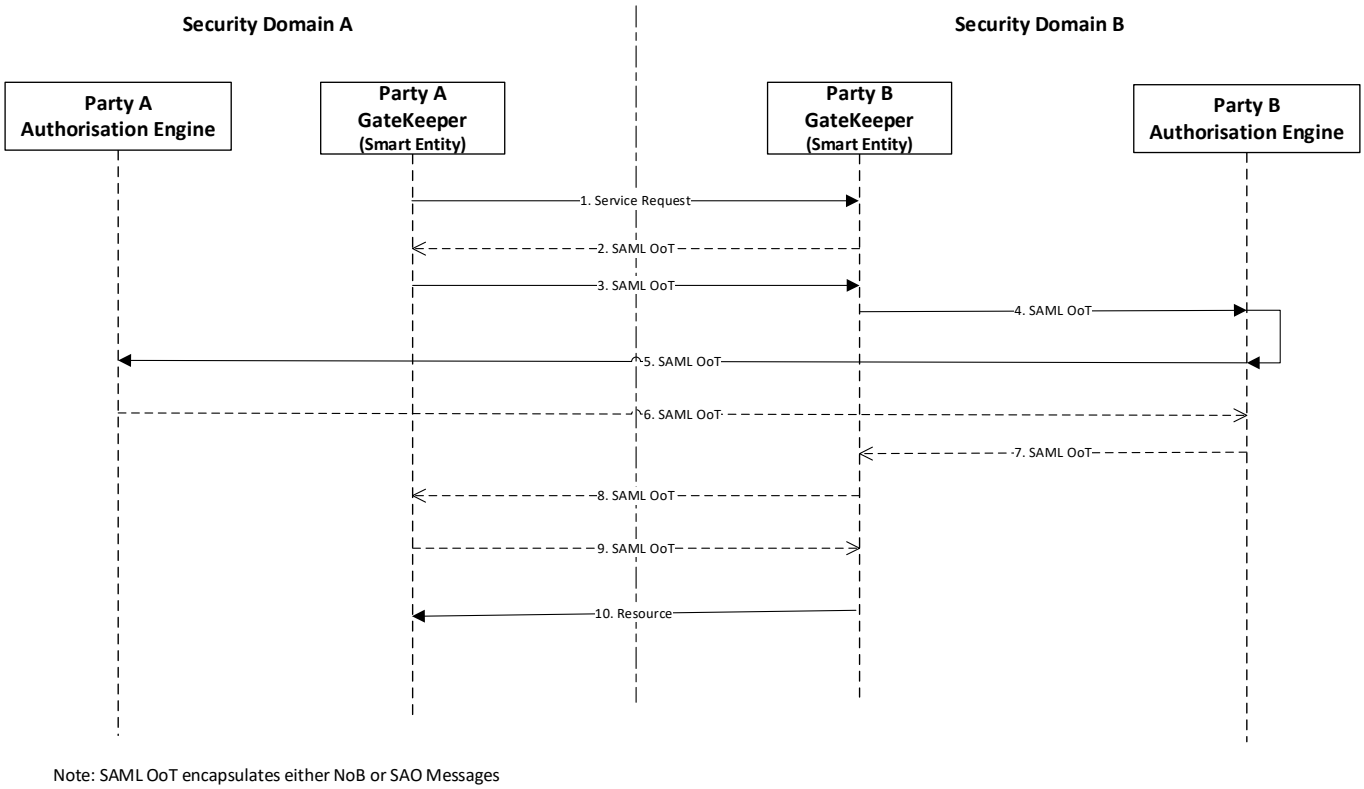


Figure 6: Access Control Conversations between Entities in Distributed Systems.

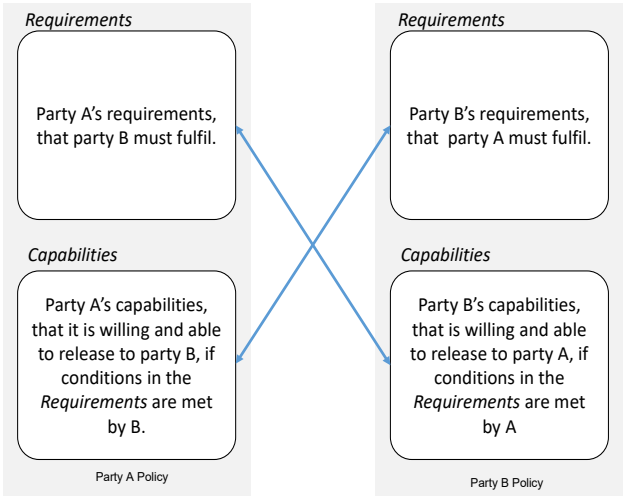


Figure 7: Obligation of Trust Policy Architecture

ii. *Capabilities* is used to express the competences (or services or features) a party is prepared to make available to another party, provided conditions expressed in its *Requirements* section are fulfilled. Thus, *Requirements* and *Capabilities* represent a policy architecture that two or more cooperating parties can leverage to assure trustworthiness, privacy and confidentiality concurrently.

The significant advantages of this policy construct include as follows:

i. It is a derivative of XAMCL and SAML standards, making it not too difficult to implement the *Matching Algorithms* and *Messaging* constructs;
ii. It is flexible to fit into any application context and has the ability to scale proportionately;
iii. It is extensible.

5. Discussions

The infrastructure presented here uses industry standards such as XACML, SAML, and frameworks including FIM, OoT for distributed access control in IoT environments. The infrastructural subunits are modular and distributed in manner adaptable to use cases where IoT computing resource are constrained. By design, it is expected that for an IoT entity to wade off application layer access intrusions, the gate keeper deny all access request by default. Consequently, to gain access to IoT resources depends on the evaluation of trusted assertion from a corresponding authorization service based on the outcome of *Matching Algorithms* by the PDP using the Policy Sets.

In implementation, there are two ways SAO can be constructed: firstly, the SAO can encapsulate digitally signed *Requirements* and *Capabilities* of a party. This signifies that this asserting party is willing and capable of providing the *Capabilities* if and only if the relying party meets the rules described in its *Requirements*. Secondly, in alternative, the SAO can comprise digitally signed *Capabilities* of the asserting party and the *Capabilities* of the relying party. In this case, it shows that the asserting party agrees to release the *Capabilities* provided the

relying party can reciprocate by releasing its own *Capabilities*.

In scenarios where incremental building of trustworthiness is required, more than one attribute would be required in the rule expressions in such manner that one of the *Subject Descriptors* of the policy indicates the initial attribute to start the first degree trust negotiation as may be required by the parties. Additionally, it can be assumed that parties may not be willing to disclose attributes of privacy value at the first round of the negotiation. In this regard, the policy specification should be such attributes required to build trust are arranged in order of less sensitive to high sensitive attributes. Alternatively, d*irect trust* between an IoT entity and its local IdP/AA, can be the basis for starting trust negotiation. The assumption here is that initial information provided by a party is insufficient to breach its privacy or undermine the confidentiality of the protected resources. This initial phase, in theory, is sufficient to counter any attempt by a malicious party to conduct probing attacks[23], usually related with trust negotiations. In this, it is further supposed that if the conversation parties decide to withdraw at the initial stage of the trust negotiation phase, their risks exposure can be significantly reduced. Moreover, if any of the parties is a hostile party, then this early interaction should filter out the access request, and terminate the conversation.

Whereas the first degree of the trustworthiness as described above is inadequate to gain access to IoT services, the parties may provide other levels of trust, which can be specified in the policy construct to help each other reach their various goals. To make this negotiation phase privacy aware, an entity can simply sends its SAML OoT containing its security *Requirements* and *Capabilities* across to the other entity. The later party, uncertain whether the other party will conform to its security settings, cannot disclose sensitive information, but correspondingly respond with another SAML OoT that describes its competences and security requirements. This iterative process operationally initializes privacy trust building and interexchange of applicable attribute information in intuitive way, which can result to a number of iterations until both parties are willing to work together.

It is obvious that the prevailing scenarios above is no way a guarantee or assurance that the parties will conform to each other's privacy, so the SAO offers a strong practical protocol that ensures conversation parties generate and interexchange digitally signed difficult-to-repudiate documents containing contextual information that can be admissible in the courts of law.

6. Conclusion

The infrastructure presented here introduces a powerful approach to identity and access management in distributed IoT environments in trust negotiation fashion. It has shown how malicious party's effort to intrude into an IoT system can be thwarted in real-time by gradual and bilateral negotiation to establish trust first before disclosure of protected resources in either direction. Privacy protection and trustworthiness, are behavioural, and possess obligatory expectations, it then implies that privacy and digital trust require a degree of assurance more than traditional security measures can provide. Our infrastructure has addressed security, privacy and safety in situations whereby IoT entities have to solve problems across multiple domains in more trustworthy, adaptive and secure manner. Equally, our approach allows both parties in conversation to mutually address their security, privacy and safety concerns as opposed to one-way unilateral protection mostly used by the party providing services.

Moreover, we have presented a novel infrastructure with distributed access control components in a fashion that access control Policy Decision Points (PDP), Identity/Attribute Authority (IAA) providers can be delegated to external trusted parties while the constrained IoT system handles context and Policy Enforcement Point (PEP).

Furthermore, by allowing parties to express their access rules and services in *Capabilities* and *Requirements* policy elements, a fine-grained access decisions can improve security and safety. Besides, addressing trust, privacy and confidentiality in a mutual way, our infrastructure provides accountability and conflict resolution approach, which are vital factors for typical IoT distributed deployments.

Conflict of Interest

The authors hereby declare no conflict of interest.

References

[1] U.M Mbanaso, G.A Chukwudebe, B. Bamidele "Holistic Security Architecture for IoT Technologies," in *13th International Conference on Electronics, Computer and Computation (ICECCO)*, 2017, pp. 11–16.

[2] U.M Mbanaso, G.A Chukwudebe "Requirement Analysis of IoT Security in Distributed Systems," 2017.

[3] U. M. Mbanaso, G. S. Cooper, D. Chadwick, and A. Anderson, "Obligations of trust for privacy and confidentiality in distributed transactions," *Internet Res.*, vol. 19, no. 2, pp. 153–173, 2009.

[4] U. M. Mbanaso, G. S. Cooper, D. W. Chadwick, and S. Proctor, "Privacy preserving trust authorization framework using XACML," *Proc. - WoWMoM 2006 2006 Int. Symp. a World Wireless, Mob. Multimed. Networks*, vol. 2006, pp. 673–678, 2006.

[5] R. Ross, M. McEvilley, and J. Carrier Oren, "Systems Security Engineering: Considerations for a Multidisciplinary Approach in the Engineering of Trustworthy Secure Systems," vol. 1, 2016.

[6] E. Number and H. Manufacturing, "Privacy and data protection."

[7] U. M. Mbanaso, G. S. Cooper, D. Chadwick, and A. Anderson, "Obligations for Privacy and Confidentiality in Distributed Transactions," *Ifip Int. Fed. Inf. Process.*, pp. 69–81, 2007.

[8] T. Ryutov, L. Zhou, C. Neuman, T. Leithead, and K. E. Seamons, "Adaptive trust negotiation and access control," *Symp. Access Control Model. Technol.*, p. 139, 2005.

[9] H. Gao, J. Yan, and Y. Mu, "Dynamic Trust Model for Federated Identity Management," *Netw. Syst. Secur. (NSS), 2010 4th Int. Conf.*, vol. 2010, pp. 55–61, 2010.

[10] A. Bhargav-Spantzel, A. Squicciarini, and E. Bertino, "Integrating Federated Digital Identity Management and Trust Negotiations," *CERIAS Tech Rep. 2005-46*, pp. 1–15, 2005.

[11] J. G. Alessandro Acquisti, "Privacy and Rationality in Individual Decision Making," *IEEE Secur. Priv.*, vol. 3, pp. 26–33, 2005.

[12] Information Commissioner's Office, "Preparing for the General Data Protection Regulation (GDPR) 12 steps to take now," *Iso*, p. 11, 2016.

[13] Information Commissioner's Office, "Overview of the General Data Protection Regulation (GDPR)," p. 43, 2017.

[14] C. Weyrich, Michael und Ebert, "Reference Architectures for the Internet of Things," 2016.

[15] P. Fremantle, "A reference architecture for the internet of things," *WSO2 White Pap.*, vol. 0, p. 21, 2014.

[16] Symantec, "An Internet of Things Reference Architecture," *Symantec White Pap.*, pp. 1–22, 2016.

[17] K. E. Skouby and P. Lynggaard, "Smart home and smart city solutions enabled by 5G, IoT, AAI and CoT services," *Proc. 2014 Int. Conf. Contemp. Comput. Informatics, IC3I 2014*, pp. 874–878, 2014.

[18] A. Torkaman and M. A. Seyyedi, "Analyzing IoT Reference Architecture Models," *Int. J. Comput. Sci. Softw. Eng. ISSN*, vol. 5, no. 8, pp. 2409–4285, 2016.

[19] S. V. Nath, "IoT architecture," *Internet Things Data Anal. Handb.*, pp. 239–249, 2017.

[20] D. Chadwick, G. Inman, and N. Klingenstein, "Authorisation using Attributes from Multiple Authorities – A Study of Requirements," p. 4.
[21] B. E. Bhargav-Spantzel Abhilasha, Squicciarini Anna Cinzia, "Identity Management Concepts Technologies Systems," *IEEE Secur. Priv.*, vol. 5, no. 2, pp. 55–63, 207AD.
[22] OpenID, "OpenID Decentralized Authentication," 2017. [Online]. Available: https://openid.net/.
[23] K. E. Seamons Winslett, M. & Yu, T., "Limiting the Disclosure of Access Control Policies during Automated Trust Negotiation," *New York Distrib. Syst. Secur. Symp.*, pp. 1–11, 2001.
[24] D. Chadwick, G. Zhao, S. Otenko, R. Laborde, L. Su, "Building a Modular Authorization Infrastructure," *Kent Acad. Repos.*, pp. 5–15, 2010.
[25] B. Parducci and H. Lockhart, "eXtensible Access Control Markup Language (XACML) Version 3.0," *OASIS Stand.*, no. January, pp. 1–154, 2013.
[26] L. S. V *et al.*, "Security and Privacy Considerations for the OASIS Security Assertion Markup," *Management*, no. August, pp. 1–33, 2004.
[27] V. Felmetsger, "Security Assertion Markup Language (SAML) SAML as OASIS Standard," 2006.
[28] U. M. Mbanaso, "Design of Obligation of Trust Protocol," no. May, pp. 19–20, 2008.
[29] A. S. Elmaghraby and M. M. Losavio, "Cyber security challenges in smart cities: Safety, security and privacy," *J. Adv. Res.*, vol. 5, no. 4, 2014.
[30] OWASP, "Internet of Things Top Ten," 2017.
[31] Gartner, "Leading the IoT, Gartner Insights on How to Lead in a Connected World," *Gart. Res.*, pp. 1–29, 2017.
[32] Information Commissioner's Offfice, "Guide to the General Data Protection Regulation (GDPR)," p. 153, 2017.
[33] U. M. Mbanaso, "Privacy Preservation Architecture for Authorization Infrastructure."
[34] K. E. S. Tatyana Ryutov, Li Zhou, Clifford Neuman, Travis Leithead, "Adaptive trust negotiation and access control," 2005, pp. 139–146.
[35] U. M. Mbanaso, G. S. Cooper, D. Chadwick, and A. Anderson, "Obligations of trust for privacy and confidentiality in distributed transactions) "Obligations of trust for privacy and confidentiality in distributed transactions Obligations of trust for privacy and confidentiality in distributed transactions," *Obligations Trust Priv. confidentiality Distrib. Trans.*, vol. 19, no. 2, pp. 153–173, 2009.
[36] U.M. Mbanaso, Centre for Cyberspace, Nasarawa State University, "Personal Data Privacy and Security - Who , What , When , Why , Where and How ?," in *DIRISA National Research Data Workshop, Pretoria South Africa*, 2018, no. June, pp. 1–8.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 4, No. 1, 50-72 (2019)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Extending the Life of Legacy Robots: MDS-Ach via x-Ach

Daniel M. Lofaro[*,1], Magdalena Bugajska[2], Donald Sofge[3]

[1]*Navy Center for Applied Research in Artificial Intelligence, U.S. Naval Research Laboratory, Washington D.C.*

[2]*Intelligent Systems Section, U.S. Naval Research Laboratory, Washington D.C.*

[3]*Distributed Autonomous Systems Group, U.S. Naval Research Laboratory, Washington D.C.*

ARTICLEINFO

Article history:
Received: 25 August, 2018
Accepted: 22 December, 2018
Online: 21 January, 2019

Keywords:
Legacy Robots
Middleware
Control Framework
Real-Time

ABSTRACT

Our work demonstrates how to use contemporary software tools on older or "legacy" robots while keeping compatibility with the original control, tools, and calibration procedures. This is done by implementing a lightweight middle-ware called MDS-Ach connected directly to the hardware communications layer of the robot's control system. The MDS-Ach middle-ware, which relies on the x-Ach methodology, was specifically designed for Xitome Mobile Dexterous Social (MDS) Robot which was released in 2008. The MDS Robot is actively used in multiple research facilities including the United States Naval Research Laboratory. This middle-ware gives the MDS Robot the bleeding edge software capabilities of today's robot by implementing the x-Ach real-time processes based computer control architecture. MDS-Ach controls the robot over its low level hardware communications interface (CAN-Bus). This communication controlled and implemented by a real-time daemon process. Controllers communicate with the real-time daemon via a ring buffer shared memory with network capabilities. The ring buffer shared memory is a "first-in-last-out" and is non-head-of-line blocking. All of the latter ensures non-blocking reading and writing of the latest data even while newer data is being added to the buffer. The UDP and TCP protocols can be implemented depending on reliability and timing requirements. Secure communication between networked controllers is implemented via tunneling over SSH if needed. The MDS-Ach middle-ware is designed to allow for simple and easy development with modern robotic tools while adding accessibility and usability to our non-hardware-focused partners. We present an implementation of real-time collision avoidance and a robust inverse kinematics solutions within the MDS-Ach system. We include detailed examples of collision avoidance, inverse kinematics implementation, and the software architecture and tools.

1 Introduction

The goal of this work is to extend the life of legacy robots that have high quality electro-mechanical hardware by allowing them to make use of modern day robotics software frameworks such as the Robot Operating System (ROS) [1], while keeping compatibility with their existing calibration and monitoring tools. The significance and the novelty of this work is in its focus on the compatibility with *legacy robotic systems* via the use of a lightweight middleware, a non-head-of-line blocking buffer design, and real-time network daemon. This document details our efforts in creating the latter for the MDS Robot. The MDS Robot by Xitome is a high degree of freedom (DOF) Mobile Dexterous Social (MDS) robot [2]. This research-grade robot originally debuted in 2008. The physical robot hardware was custom built to support human-robot interaction research. This includes the two 6-DOF arms each with a 7-DOF hand. Each hand has four underactuated fingers. The robot also includes a 4-DOF neck and a 17-DOF face. All of the latter items enable different facial expressions, physical gestures, and grasping within a single robotic platform.

Given the goals of this work, our system must meet

[*]Corresponding Author: Daniel M. Lofaro, U.S. Naval Research Laboratory, Washington D.C., USA - daniel.lofaro@nrl.navy.mil

https://dx.doi.org/10.25046/aj040107

Table 1: Summary of existing software robotic frameworks [4]

Framework	Concurrency Model	Data Sharing	Focus
ROS [1]	Processes	Proprietary TCP	Mobile Robots, Vision, and AI
ROS 2.0 [3]	Processes	Selectable DDS	ROS 1.0 + Real-Time Control
OpenRDK [4]	Threads	Shared memory, Proprietary TCP/UDP	Mobile Robots
Player [5]	Threads	Client/Server over TCP	Mobile Robots
YARP [6]	Processes	shared memory, proprietary TCP	Mobile & Serial Kinematic Chain Robots
MARIE [7]	Processes	Many (3rd Party)	Connecting Different Frameworks
OpenRTM-aist [8]	Threads	CORBA	General Robotics
Orca [9]	Processes	ICE[10]	Mobile Robots
JAUS [11]	Processes	Proprietary TCP/UDP	Standardization, Unmanned Systems
x-Ach [12]	Processes	Shared Memory, Proprietary TCP/UDP	Real-time Robotics

specific requirements. Firstly, the well-tuned and refined calibration procedures that are extensively documented for the robot must remain unchanged while extending the robot's capabilities. Similarly, existing monitoring tools must remain functional, but often cannot be modified. Furthermore, with robotics and Artificial Intelligence (AI) research back in the spotlight and becoming more mainstream in recent years, a robot system must implement additional safety controls to allow researchers with no mechanical or electrical engineering background to use the robot while keeping the risk of damage to the physical system low. The system must also be capable of working securely over a network. Finally, the system must not be confined to a single programming language, allowing the user to use "the right tool for the right job."

To keep compatibility with the legacy software, direct control is implemented by commanding the MDS Robot over the Controller Area Network (CAN) bus via a dedicated real-time daemon. A process based controller approach is used for the control system using a "first-in-last-out" (FILO) non-head-of-line blocking ring buffer type of shared memory. This allows for the use of multiple programming languages in the same system. This also gives each controller the ability to *read the newest data first* which is typically of most importance to real-world robot controllers. SSH tunneling is used when secure connections between network connected controllers is required. Real-time collision avoidance with a robust inverse kinematics solution is implemented within the MDS-Ach system allowing for a multitude of types of users to safely control the robot. In the following sections, we describe our methodology and provide usage examples of the system. We also include, in the appendix, a description of the tools and the implemented API. Please note that this paper is heavily based on and is an addition to the authors previous work in Lofaro et. al. [13]. This work adds detailed instructions and examples of how to utilize the MDS-Ach system.

2 Background

There are currently many implementations of middleware for robot systems. The most notable one is the Robot Operating System which is commonly known as ROS[1]. ROS is a middle-ware which is TCP-based. It allows for communication between different controllers other over "topics." This is typically done to connect the hardware systems of the robot such as the sensors and actuators to the logic and control. Its biggest strength is large ecosystem. Additionally, ROS is most useful for systems that require many controllers but do not require real-time capabilities. Currently ROS 2.0 [3] is being developed. ROS 2.0 will add real-time capability to the system. Because of the latter MDS-Ach is written specifically to be compatible with both ROS 1.0 and 2.0.

The YARP system, or Yet Another Robot Platform, is a C/C++ based middle-ware. The purpose of YARP is to connect control processes, sensors, and actuators. YARP is tested on Windows, Linux, and OSX [6]. It uses shared memory, over TCP when needed, for communication between the YARP server and clients. Non-blocking and latest data first reading is implemented via double and triple buffers. YARP is currently limited to a C/C++ API.

OpenRDK is an open source framework for robotics [4]. It uses socket communication and shared memory to implement its thread based architecture. This impressive control system utilized linking techniques and blackboard-based communication to allow for input/output data port conceptual system design. OpenRDK is a thread based design. Our desired system is processed based not thread based.

Joint Architecture for Unmanned Systems, also known as JAUS [11], was originally an initiative started in 1998 by the United States Department of Defense (DoD) to develop an open architecture for the domain of unmanned systems. JAUS was formerly known as Joint Architecture for Unmanned Ground Systems (JAUGS) and is built on five principles: vehicle platform independence, mission isolation, computer hardware independence, technology independence, and operator use independence. Still in use by the DoD, JAUS communicates with other systems over TCP and/or UDP. Though a formidable system, the public ecosystem is relatively small when compared to competitors. A comparison of the above-mentioned middleware, but also other robotic frameworks can be seen in Table 1.

The *x-Ach* system is based on the Ach IPC, or inter process communication [12]. Current implementation

of *x-Ach* include Hubo-Ach [12] for the Hubo series of robots, Android-Ach for phones, Shoko-Ach [14] for the underwater legged robot AquaShoko, MDS-Ach (this work) for the MDS Robot, and more. The *x-Ach* system is lightweight with non-head-of-line blocking like OpenRDK, however it uses processes for each controller instead of threads. Like YARP, we use the idea of newest data first. Like OpenRDK, we use shared memory and offer a choice between TCP and UDP depending on need. *x-Ach* is compatible with multiple languages including C/C++, Python, Java, etc. ROS currently has the largest robot controller ecosystem, however it is not real-time; ROS 2.0 will be real-time when completed. To leverage ROS' ecosystem, *x-Ach* is written to be compatible and easily integrated with either ROS versions.

3 Methodology

This section details the methodology and implementation of the MDS-Ach system and is based on our original paper [13].

3.1 Controller Area Network Communication

A key goal of this work is to keep compatibility with older robot's legacy software and well-defined calibration procedures. In this case the specific robto in question is the MDS Robot. We created the MDS-Ach system to connect directly to the robot via the CAN (Controller Area Network) bus. This is the communications bus that is concurrently used to control each actuator on the MDS Robot and is also used with the legacy system, in this case the MDS Motion Server. The CAN bus is specifically designed designed for multiple devices/controllers to communicate over it. The latter allows all of the legacy software, utilities, and tools to run, monitor, and calibrate the robot while allowing for integration with the state of the art robotic software. The direct communicating via the CAN bus results in keeping compatibility with the original software and tools without having to modify, recompiled, or in any way adjust these tools in order for the MDS-Ach system to run.

3.2 x-Ach

The *x-Ach* system has a process based architecture. This means that it runs individual controllers as independent, synchronous and/or asynchronous, processes. Each process communicates with each other over the IPC called Ach which is a circular buffer and is non-head-of-line-blocking [12].

Ach was chosen because it is low-latency (key for real-time control) and has a *first-in-last-out* (FILO) buffer. This allows controllers to get the newest information first while retaining the ability of reading older information at a later time if needed. The latter is very important to real-time robotic systems. Packed c-structs are used for messages types in order to keep the system archatecture agnostic. This means that x-Ach controllers running on different platforms (i.e. x86, amd64, ARM, and other systems) can communicate with each other despite the different memory block sizes. It is important to note that all of the memory types are well defined within the packed c-struct. For example a (int32_t) is used instead of (int) to ensure data congruence between different architecture.

Controllers communicate with each other over Ach channels. Each controller has a standardized input channel called "reference" (*ref*), a standardized processed reference output channel called "processed reference" (*p-ref*), and a standardized output channel called "state" (*state*). Details on the reference (*ref*) channels can be found in (Section 3.2.1) and details on the state (*state*) channel can be found in (Section 3.2.2).

3.2.1 Reference Channels

Each process based controller has two reference channels. One is a standardized input channel called "reference" (*ref*) and one is a standardized processed reference output channel called "processed reference" (*p-ref*) The latter two channels are where other controllers can write, or publish, to and read from, or subscribe to, respectively. It is important to note that the reference (*ref*) channel does not bind. This means that multiple controllers can write, or publish, simultaneously to the same reference channel. The controller will only use the most recent message received and only utilize the other older messages if it is specifically needed.

The process can synchronize its control loop to the incoming reference (*ref*) input or be asynchronous (i.e. free running). The latter is useful for use and development of synchronous and asynchronous controller systems and controllers. Both of the reference channels, the input reference (*ref*) and the process output reference (*p-ref*), are identical in structure.

As stated in Section 3.2, the reference channels consist of packed c-structs. The size of the structure is dependent on the number of degrees of freedom (DOF) of the robot of the given robot. The reference structure contains joint-space references (*JS-ref*) and work-space (Cartesian space) references (*WS-ref*) (see Figure 5). This is in the same structure to allow for less complexity in the controllers' number and types of sources and sinks.

3.2.2 State Channel

Each MDS-Ach controller has one channel for the *state* what is real-only by other processes. Other controllers can read the most up to date state of the robot/controller by reading the *state* channel even if the given robot/controller is currently updating the state. For the implimentaiton on the MDS a daemon is created called the *MDS-Ach daemon* (see Section 3.3). This daemon publishes the most recent joint-space states including, but not limited to, actual position of the

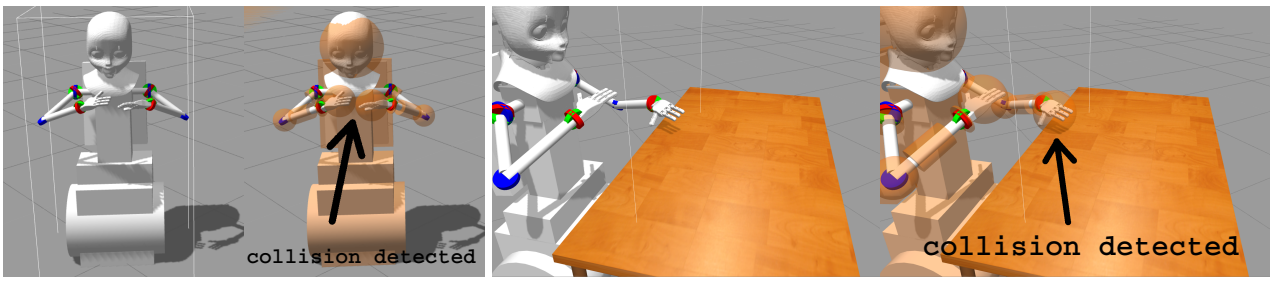


Figure 2: High-resolution MDS Robot model **(LEFT)** next to the high-resolution MDS Robot model with the low-resolution collision models overlapped **(CENTER LEFT)**. The collision models are simplifications of the high resolution geometry and are denoted in orange. The collision model is "over-sized" in order to help the system detect a collision before an impact actually occurs. *Self Collision Avoidance Test*: Robot does not collide when each hand is told to go to the same location. **(RIGHT)** *World Collision Avoidance Test*: Robot does not collide when the hand is commanded to a position that will collide with an object in its workspace. In this example the hand stops before it hits the table.

joint, last received reference/commanded position, joint load/current, etc. Like the reference channels the state channel is a packed c-struct. The size of the structure is dependent on the number of DOFs of the robot. An example use case for the state channel is the collision detection daemon in Section 3.4.2. This daemon reads the *state* channel and applies those values to its internal model to check for collisions in real-time. The daemon then outputs collision state of the robot on its state channel. This is done without disrupting the real-time performance of the *MDS-Ach Daemon*.

3.3 Daemon

The MDS-Ach daemon is the bridge between the CAN bus, which commands the MDS Robot, and the process based x-Ach controllers. The goal of all x-Ach daemons, in cluding the MDS-Ach daemon, is to be the "*driver*" for the given robot. The CAN bus is half-duplex running at a rate of 1.0 $Mbps$. Joint-space control and sensor feedback (i.e. reference and state information) is sent over the CAN bus to and from the daemon. The MDS-Daemon is specifically calibrated to keep the load on the CAN bus at approximately 60% of its bandwidth saturation. The latter is done to help guarantee real-time performance and on-time data delivery. The x-Ach daemons have been implemented and tested on multiple different types of robots with different CAN packet structures and a varying amount of communication buses. 200 hz real-time performance was achieved when utilizing multiple CAN buses (two) along with a specialized packet structure and the the utilization of the PREEMPT_RT linux kernel [15] as seen in our previous work [12]. The MDS requires the use of only one CAN bus with a BAUD rate of $1Mbps$ and state-full packet structure. All of the latter limitations require us to run at 10 hz to guarantee real-time performance, with sub ms accuracy, for this specific robot.

The MDS-Ach daemon has an optional first-order real-time joint-space position smoothing filter. This filter is applied to the input from the reference (*ref*) channel (Section 3.2.1) before being applied to the control and sent to the robot over the CAN bus. The first-order real-time joint-space position smoothing filter converges to within 95% of the reference input within 4.0 seconds and is enabled by default. This filter was added in order to reduce acceleration and jerk of each joint without limiting the maximum velocity.

The MDS-Ach daemon reads the reference command (as described in Section 3.2.1) in real-time and sends the command data over the CAN bus. This process is asynchronous in reference to the robot's actuators control loop. If multiple commands are sent within one cycle only the newest one is read and processed. There is a zero order hold is there are no new commands between given control cycles. The except to this is during setup phases such as "homing", resetting/error correcting, and other hardware setup/configuration specific commands. Additionally, during each cycle the MDS-Ach daemon requests and reads the information from the sensors via the CAN bus and writes it to the *state* channel (see Section 3.2.2).

3.4 Input Pipeline

The Main Controller receives the user-level commands in either joint-space or Cartesian-space, over the ACH shared memory. The latter input goes through the pipeline described in section below before it is sent to the Daemon (Section 3.3).

3.4.1 Joint Mux

The Joint Mux is an event-based process that takes in joint-space references from multiple processes. It updates the primary joint-space reference, which is sent to the MDS-Ach daemon, only with the joint-space references that are controlled by and updated by a given controller while preserving the ones that they controller does not have write permission to and/or were not modified. The MDS-Ach daemon is then sent the resulting consolidated (muxed) reference command. The purpose of this is to allows multiple controllers to update individual joints without conflicting with (overwriting) other joint commands while preserving a common message type.

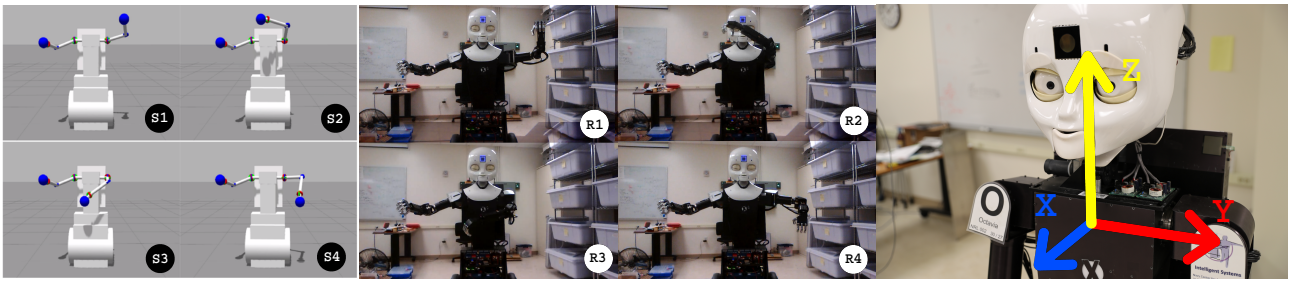


Figure 3: Example of the MDS-Robot (MIDDLE) and its simulation (LEFT) utilizing the ik solver, joint smoother, and self-collision avoidance to move its hand in a square pattern while keeping all rotational degrees of freedom in the null space. (RIGHT) Coordinate system for the MDS Robot when using MDS-Ach.

3.4.2 Real-Time Collision Detection

The fully integrated real-time collision detection system utilizes the opensource simulator GazeboSim and ODE [16, 17] to ensure safe system operation. To guarantee the real-time performance, the system must detect self-collisions as well as collisions with objects in the environment within one time step of the MDS-Ach daemon (as defined in 3.3). To maintain compliance with the real-time deadline unneeded features of the simulator, such as *physics*, are disabled. This results in a reduced computational load. Additionally a geometrically simplified/reduced model of the MDS robot is created. This model is comprised of basic shapes such as cylinders, boxes, and spheres. This simplified model is used to create the collision model to further reduce computational complexity and improve performance of the system. The performance improves because detecting collisions between two spheres only requires comparing the euclidean distance between the spheres centers and to the sum of the spheres' radii. There is no collision as long as the sum of the radii is less than the euclidean distance from the centers. The same can be said for cylinders if the conditional measurement is made from the center of the sphere to the closest point on the axial line of the cylinder. Figure 2 shows the MDS Robot model with the collision model overlapped.

The collision state is written in state the message format, as described in Section 3.2.2, to the *Main Controller*. If the reference position is free of collisions, a joint-space reference is sent to the the MDS-Ach daemon via the Joint Mux. The MDS-Ach daemon will execute the motion on the physical robot. If a collision is detected, the most recent safe joint-space reference values will be used for all the joints of the affected arm(s). "Expert" user is able to overwrite this behavior, if needed, by sending a joint-space command from the main controller.

3.4.3 Joint-Space Smoothing

When the robot is operating in situations where it is moving between joint-space configurations or required to stop its motion to avoid collision/self-collision (see Section 3.4.2), the safety of the robot's joints needs to be ensured. Torque due to high joint-space acceleration is one of the primary causes of robot joint damage. We need to reduce the torque applied to the joints without causing joint-space overshoot. Furthermore this needs to be done in real-time and on-line. The motion at this level can not be pre-planned so it can be used in real-time tasks such as servoing or world interaction. This section shows how we reduce the acceleration, which reduces torque, on the joints while reducing the overshoot. The latter is done by applying the filter shown in (1).

$$\theta_n = \frac{(\theta_{n-1}L - 1) + \theta_{des}}{L} \tag{1}$$

Where θ_n is the output of the filter which is the new reference position (angle) the joint will be commanded to at step n; θ_{n-1} is the reference position to the joint from the previous time step, i.e. $n-1$; L is the weight of the filter (defined by its integer length); and θ_{des} is the desired reference position in joint-space that the joint is requested to go at time step n.

The position as recorded from the joints' encoders (θ_{enc}) are used to add joint-space compliance to the system. This is done by replacing θ_{n-1} in (1) with θ_{enc} as seen in (2). The use of the measured angle allows us to take advantage of the natural compliance in the system and magnify it. When the filter is applied it results in a pose "sag" due to gravity.

$$\theta_n = \frac{(\theta_{enc}L - 1) + \theta_{des}}{L} \tag{2}$$

3.4.4 Inverse Kinematics

We utilize the "Inverse Jacobian" method for our online inverse kinematics solver. This is used to the joint-space angles for the commanded Cartesian-space (work-space) positions [18, 19]. The joint limits are utilized when constructing the the Jacobian during each cycle. During each step of the search the resulting pose is checked for self-collisions via ODE (see Section 3.4.2). A new *intermediate* goal position is created if a collision is found. The process is then repeated. The joint-space configurations found from the Jacobian IK is then passed through the filter discribed in (1) to prevent large steps in joint-space. Each step of the resulting filtered positions are then checked in real-time by the collision checker (Section 3.4.2) before being sent to the MDS-Ach daemon.

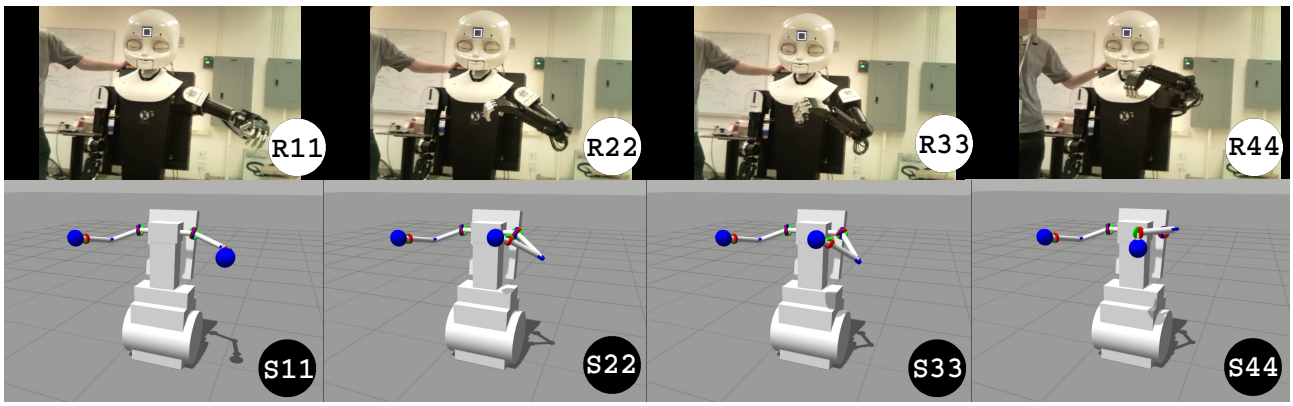


Figure 4: Example of the MDS-Robot (TOP) and its simulation (BOTTOM) utilizing the IK solver, joint smoother, and self collision avoidance to move its hand in a box with the coordinates $(x, y, z, \theta_y) = (0.35m, 0.45m, 0.25m, null) \rightarrow (0.35m, 0.10m, -0.05m, null) \rightarrow (0.45m, 0.10m, -0.05m, null) \rightarrow (0.45m, null, null, 90^o)$ while keeping some linear and some rotational degrees of freedom in null space.

The Inverse Jacobian IK method is used because it allows control the required work-space degrees of freedom while the other degrees of freedom remain in *null space*. The MDS Robot is used for world interaction tasks such as opening doors (via pushing), grabbing cups, pushing buttons etc. thus having work-space degrees of freedom in *null space* is needed. In the door opening example (via pushing) the robot only degree of freedom that strictly matters is the x value (out of the robot's chest - see Figure 3). The x value defines the distance the robot pushes the door open. The height (z) and the left/right distance (y) does not matter as much and thus can be left in null space. Furthermore the orientation (all three degrees of freedom) of the hand also belongs in the *null space*.

Solving for the required degrees of freedom for the given task, and leaving all the others in the *null space*, allows less iterations/computations for the solver resulting in a faster solving time and the ability to run in real-time. For MDS Robot, real-time constraints are satisfied if the joint-space values for a given Cartesian-space (joint-space) position are calculated in less than 0.5 *sec*. This is achieved when the desired Cartesian-space position can be described with DOF $\leq$4. The higher order of the required position requirements (i.e. the less degrees of freedom in the *null space*) the more time (on average) is required to solve the joint-space solution. A measured $\sim$ 0.7 *sec* is required to solve for a 5-DOF position and $\sim$ 1.3 *sec* for a 6-DOF on our contemporary computers using a single core. It is important to note that the latter calculated times are averages, also a solution is not guaranteed to be found even if one exists.

Figure 3 shows the MDS-Robot and its simulation utilizing the inverse kinematics daemon. The figure shows the MDS-Robot commanding its end-effector to move in a box pattern in real-time. A detailed description of the test can be found in Section 4.1.

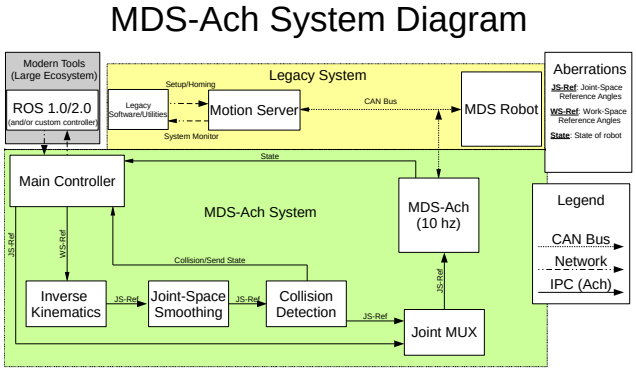


Figure 5: MDS-Ach middleware: A real-time process based control system used to extend the research life of the MDS Robot.

3.4.5 Network Daemon

The network daemon utilizes *achd*, as a part of the Ach library, to share data over a "socket" supported network. The *achd* daemon pushes the state data from the primary control computer on the robot to the external "user-level" computer controller when new state data is available. When a reference command is updated by the user-level computer controller it is pushed to the primary control computer over the network by the *achd* daemon. Updates are only pushed or pulled when there is new state or reference data in order to conserve bandwidth. UDP is used by default for data transport to help facilitate tighter real-time performance. TCP or TCP over a SSH tunnel can be used by the user depending on timing, reliability, and security requirements. More details can be found in our previous work [12].

3.4.6 Security

Intercepting and spoofing network traffic is a major threat to any robot system and the threat persists even on a properly configured Unix/Linux computer. These "man-in-the-middle" attacks allow a third party to control the robot using the existing controllers if they can

spoof the feedback data [20]. To prevent a man-in-the-middle attack over the network, a secure SSH tunnel with pre-shared keys is used to transmit and receive state and reference data between the main controller and the user-level controller. Since SSH tunnel operates over TCP, the network throughput between the user-level controllers and the Main Controller is reduced, and may even affect the real-time performance.

3.5 External Framework Bridge

The ROS 1.0 and ROS 2.0 bridge allows us to make use of the extensive ROS ecosystem and serves as proof of concept for extensibility of this middleware. This bridge talks directly with the Main Controller via the Ach shared memory or the network daemon (3.4.5). This controller converts the state data published by the MDS-Ach system into ROS messages to be published again on specified ROS topics. To reduce latency (1) the state topic of ROS is synchronous with the state channel of MDS-Ach, and (2) the reference channel is synchronous with the state topic of ROS (seen Fig. 5).

4 Testing

At the conclusion of the MDS-Ach middleware development, we tested the software to confirm its functionality. We present the testing procedures and results of the inverse kinematics system (Section 4.1), the *live* self-collision detection system (Section 4.2), and the *live* external-collision detection system (Section 4.3).

4.1 Inverse Kinematics Test

To validate the usability of the inverse kinematics system, we performed couple tests to show its ability to move the end-effector to a position (4.1.1) and an orientation (4.1.2) in the Cartesian-space (work-space) while keeping other variables in the null-space. The coordinate system used for the IK tests is shown in Figure 3; the origin is the orthogonal projection of the base of the neck onto the line defined by the rotations points of the shoulders. x-y plane is horizontal. x-z plane extends in front and behind the robot while y-z plane extends to the sides. All signs of rotation follow the "right hand rule."

4.1.1 Position Control IK Test

This section presents results showing that the end-effector can be moved between multiple linear Cartesian-space (work-space) locations (x,y,z) without constraining the end-effector orientation. The left arm is commanded and moved to the locations shown in (3) where S1, S2, S3, and S4 are the coordinates for the simulated robot and R1, R2, R3 and R4 are the coordinates for the real robot.

$$\begin{aligned} S1 = R1 &= (0.40m, 0.45m, 0.25m) \\ S2 = R2 &= (0.40m, -0.05m, 0.25m) \\ S3 = R3 &= (0.40m, -0.05m, -0.25m) \\ S4 = R4 &= (0.40m, 0.45m, -0.25m) \end{aligned} \tag{3}$$

The screenshots of the resulting motion can be found in Fig. 3.

4.1.2 Orientation Control IK Test

This section presents results showing that the end-effector can be moved between multiple Cartesian-space poses (x,y,z,θ_y) while keeping the remaining orientations in the null space. The left arm is commanded and moved to the locations shown in (3) where S11, S22, S33, and S44 are the coordinates for the simulated robot and R11, R22, R33 and R44 are the coordinates for the real robot. In the case of the last motion (R44 and S44), θ_y is set to 90° and a desired x to a value of $0.45m$ while all other degrees of freedom remain in the null space. The full set of coordinates for this test can be found in (4).

$$\begin{aligned} S11 = R11 &= (0.35m, 0.45m, 0.25m, null) \\ S22 = R22 &= (0.35m, 0.10m, -0.05m, null) \\ S33 = R33 &= (0.45m, 0.10m, -0.05m, null) \\ S44 = R44 &= (0.45m, null, null, 90^o) \end{aligned} \tag{4}$$

This shows that the IK process can solve for some degrees of freedom while keeping the others in null space (see Figure 4). This improves system performance when higher fidelity solutions are not required for system operation.

4.2 Self-Collision Detection and Avoidance Test

To test the self-collision system as described in Section 3.4.2, we drove the hands to the same location using the IK system described in Section 3.4.4. Multiple instances of this test were performed. In all the runs, the hands stopped before colliding. Figure 2 shows one example of the multiple self-collision tests. In this instance of the test, both the left and the right hands were told to go to the (x,y,z) coordinates $(0.3m, 0.2m, 0.0m)$. As expected, the hands did stop when the two parts of the collision model touched as shown in Fig. 2.

4.3 World-Collision Detection and Avoidance Test

To test the world-collision avoidance system described in 3.4.2, we drove the hand from the position stated in Section 4.2 out to an x value of $0.5m$ towards the table. We placed a model of a table in the simulated world where the real table would be. From there we drove the hand down in z. Similar to the results in 4.2 the hand would not move farther than the collision point between the collision model of the hand and the collision model of the table (see Figure 2). This was done for multiple objects and multiple collision locations.

5 Usage

This section documents how to start, stop, and use different MDS-Ach utilities. The MDS-Ach controls the MDS Robot via the CAN bus. It provides smoothing/filtering, multi-process control architecture, multi-language support, inverse kinematics for right and left arm, and more.

5.1 Daemon Control: mds-ach

The MDS-Ach daemon is the process running the background that controls communications between the physical (or simulated) robot and the controllers. The daemon controls the CAN bus when connected to the physical robot. Section 5.1 describes the different input options and modes of the MDS-Ach daemon. A full description of the options for the console input for the MDS-Ach daemon can be found in Appendix A.

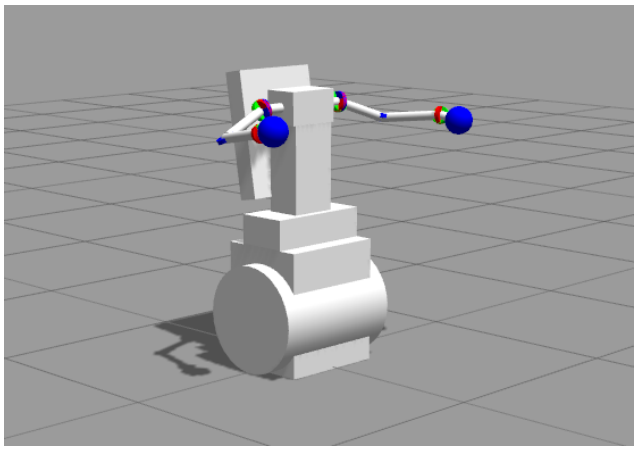


Figure 6: MDS Robot simulated in Gazebo.

5.2 Startup Procedure

This section documents the startup procedure for the MDS-Ach system.

5.2.1 Initialize

We preserved the original MDS setup procedure provided by MDS manufacturer, Xitome. Once setup is complete, which include a fairly involved homing procedure for all the joints, the user is able to keep the Xitome subsystem running. User may **not** issue commands to any of the joints using the Xitome system once MDS-Ach is started.

5.2.2 Start Daemon

It is important that the MDS robot is in its homed configuration when starting the MDS-Ach system. This is because MDS-Ach system assumes the robot's starting position to initialize the IK and the self-collision systems. If the robot is not in its "homed" position, than some joints will get a step input to go to the starting position. All starting positions are read from the **anatomy.xml** configuration file located in **/etc/mds-ach**. This file is identical to that used by the Xitome system for configuration.

The MDS-Ach system can run on the robot (Option 1) and on a simulated robot (Option 2). Option 1 sends all of the commands over the CAN bus while Option 2 sends all commands to GazeboSim[16]. Both Option 1 and Option 2 have an x-Ach[13] abstraction layer between their respective communication buses and the controller. A full description of the implementation procedure can be found in Appendix B.

5.3 Examples

Examples of how to do basic operations using the MDS-Ach system on both the real robot and the simulator can be found in Appendix C. The examples show how to start the MDS-Ach daemon on the physical and simulated robot. An image of the simulated robot can be seen in Figure 6. All example code/software can be found in Lofaro et. al. [21].

6 Utilities

This section describes the utilities available for use with the MDS-Ach system. All utilities work seamlessly with the physical and simulated robot.

6.1 MDS-Ach Console

The MDS-Console utility allows the user to read and set individual joint angle values via the command-line. It also allows the user to read the Caretesian-space pose (6 DOF) of the end-effector and set the desired target pose (3, 4, 5, and 6 DOF). The console provides control interface for all the joints listed in Table 2. We included examples of single joint interactions, but also end-effector control, in Appendix D.

Table 2: Joint abbreviations (short and long) with definitions

Definition	< *joint* > (short option)	< *joint* > (long option)
Right Elbow Pitch	REP	RightElbowFlex
Right Shoulder Yaw	RSY	RightUpperArmRoll
Right Shoulder Roll	RSR	RightShoulderAbd
Right Shoulder Pitch	RSP	RightShoulderExt
Left Elbow Pitch	LEP	LeftElbowFlex
Left Shoulder Yaw	LSY	LeftUpperArmRoll
Left Shoulder Roll	LSR	LeftShoulderAbd
Left Shoulder Pitch	LSP	LeftShoulderExt
Right Wrist Roll	RWR	RightWristFlex
Right Wrist Yaw	RWY	RightWristRoll
Left Wrist Roll	LWR	LeftWristFlex
Left Wrist Yaw	LWY	LeftWristRoll
Torso Yaw	WST	TorsoPan
Neck Roll	NKR	HeadRoll
Neck Pitch (lower)	NKP1	HeadPitch
Neck Pitch (upper)	NKP2	NeckPitch
Neck Yaw	NKY	HeadPan

6.2 MDS-Ach Read

The MDS-Read utility allows the user to view the joint space references, state, and address of each active joint. This utility updates the information at $\sim 20Hz$. It may be started or stopped at any time without affecting the overall system.

6.2.1 Prerequisites

Since MDS-Read is a monitoring tool, it must be started after the MDS-Ach is already running.

6.2.2 Startup

To start the MDS-Read utility, run the command below. The expected terminal can be seen in Figure 7.

Bash/MDS-Ach Console:

```
$ mds-ach read
```

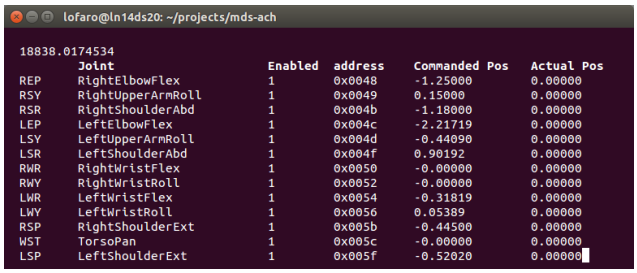


Figure 7: Expected window for ($ mds-ach read)

- **Note 1:** If you are running with the simulator the state (Actual pos) will always be zero. When running on the physical robot the real position (in radians) will be shown.
- **Note 2:** If you make more joints "active" in the **/etc/mds-ach/configs/mdsach.xml** configuration file they will automatically show up in MDS-Read.

6.3 Software Interface

The MDS-Ach system currently works with the C/C++ and Python programming languages. Appendix E describes the required libraries for each of the latter languages.

7 Operating in Joint-Space

This section documents how to use MDS-Ach to control the robot in joint-space. To move one, or multiple joint/joints the following steps must be followed:

- Open Ach Channels - This is the ring buffer shared memory that this controller communicates with MDS-Ach with.
- Create Required Data Structures - Standardized (packed) data structure that contains all of the reference and state data for the robot.
- Get Joint ID - Determine the IDs of the joint/joints to be controlled (See Table 2).
- Queue New Motor Position - Put the desired motor positions in the reference structure at the index position defined by the Joint IDs.
- Set Motor Position (put) - Write the reference structure that has been updated with the desired position for the desired joint to the reference Ach channel.
- Close Ach Channels on Exit - Though not required it is good practice to close the unused Ach channels upon exiting the controller.

Figure 8 shows the expected simulator window when following the above steps and setting the Left Elbow Pitch (LEP) and the Right Shoulder Pitch (RSP) to -0.2 *rad* and 0.1 *rad* respectively.

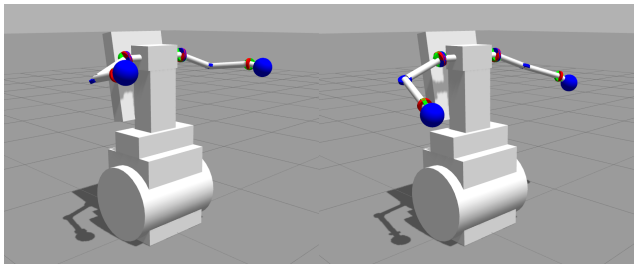


Figure 8: Expected simulator window for ($ ./mds-simple-demo). (LEFT) Before running. (RIGHT) After running.

It is important to note that Appendix F is an depth explanation and example of how to setup the Ach channels for communication with the MDS-Ach system and how to control the robot in joint-space while using the smoothing filter process. Examples are given in C/C++ and Python.

8 Operating in Cartesian-Space

In this section we discuss end-effector operations in Cartesian-space which utilize forward and inverse kinematics controllers of the MDS robot.

8.1 Built-in IK Controller (Python)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system begins to solve the IK equations automatically, when you post a new desired position on the IK reference channel. Below is a step by step of how you do this using Python.

8.1.1 Python imports

The following are required imports for the MDS-Ach system while using python: mds_ach, ach, and mds. Other imports are for the given controller implementation.

Python:

```
import mds_ach as mds
import ach
import time
import sys
import os
import math
```

8.1.2 Create Open Ach Channel for IK

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that ach.Channel() takes a string as an input. Here we opened one channel. This is a different channel then found in previous sections because it is only for the IK controller.

Python:

```
# Open ACH Channel for IK
k = ach.Channel(mds.MDS_CHAN_IK_NAME)
```

8.1.3 Make IK Structure

Similar to controlling the robot in joint-space you need to set a reference structure to the desired work-space position. For this you need to initialize the structure. See below for the initialization of the work-space structure.

Python:

```
# Make new IK structure
ikc = mds.MDS_IK()
```

8.1.4 Setting the DOF controlled

When using the built-in IK controller you need to set the number of DOF that you are controlling. With this controller you are required to set the DOF in the following order: $p_x, p_y, p_z, \theta_x, \theta_y, \theta_z$ where p_n is the position on axis n and θ_n is the rotation about axis n. For example if you DOF is set to 4 you are controlling p_x, p_y, p_z, and θ_x. If you are controlling 2 you will control p_x and p_y. In the example below we are controlling 3, i.e. p_x, p_y, and p_z. This order can also be seen in Table 5.

Python:

```
# Set the amount of DOF you want to control
dof = 3
```

We set these to the values in Table 3.

Table 3: Inverse Kinematic Values Set for Example 8.1.4

Param #	Definition	Abbreviation	Value (rad)
1	Position in X	p_x	0.3
2	Position in Y	p_y	0.2
3	Position in Z	p_z	0.0
4	Rotation in X	θ_x	Null Space
5	Rotation in Y	θ_y	Null Space
6	Rotation in Z	θ_z	Null Space

θ_x, θ_y, and θ_z are in the Null Space because we do not care where they are as long as the first three parameters are met. We can set these values to what ever we want and they will be ignored. In this case we set them to zero.

Python:

```
# Set values for work-space in
# [x, y, z, rx, ry, rz] order
eff = [0.3, 0.2, 0.0, 0.0, 0.0, 0.0]
```

8.1.5 Choosing Arm for IK

Here we pick the arm for the IK. Our options are set as enums in the mds_ach.py and mds.h includes. All options for right and left arms can be found in Table 4.

Table 4: Definitions for left and right arms using mds_ach.py in Python

Arm	Python Definition
Left	mds.LEFT
Right	mds.RIGHT

Here we set the arm to the left arm.

Python:

```
# set arm
armi = mds.LEFT
```

8.1.6 Set IK Structure

Just as in the joint-space method, we need to the values in our structure before we send it to the robot. Here we set all of the parameters from above to the structure ikc that we created.

Python:

```
# Put setting into ik structure
ikc.move = armi
ikc.arm[armi].ik_method = dof
ikc.arm[armi].t_x = eff[0]
ikc.arm[armi].t_y = eff[1]
ikc.arm[armi].t_z = eff[2]
ikc.arm[armi].r_x = eff[3]
ikc.arm[armi].r_y = eff[4]
ikc.arm[armi].r_z = eff[5]
```

8.1.7 Command the Robot

Just as with the joint-space controller you need to "put" the data on the ACH channel. Unlike the joint-space controller it will not move as soon as you send it. The controller will first have to solve the IK. Upon finding the solution the controller will send it to the robot. If a solution is found it will take anywhere between 0.1 and 5.0 seconds. If there is no solution found the robot will not move. Note: the robot limits its self to 1000 search iterations for an IK solution.

Python:

```
# put on to ACH channel
k.put(ikc)
```

8.1.8 Running the Code

To run the code do the following within the **mds-ach/examples** folder.

Bash:

```
$ python ms_ik_one.py
```

The expected terminal out can be seen in Figure 9.

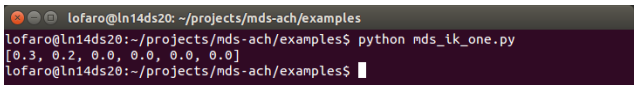


Figure 9: Expected window for ($ python ms_ik_one.py)

To check that the IK worked you can run the FK in the MDS-Ach console and/or run the simulator. The before and after of the MDS-Ach console is found in Figure 10 and Figure 11, respectively.

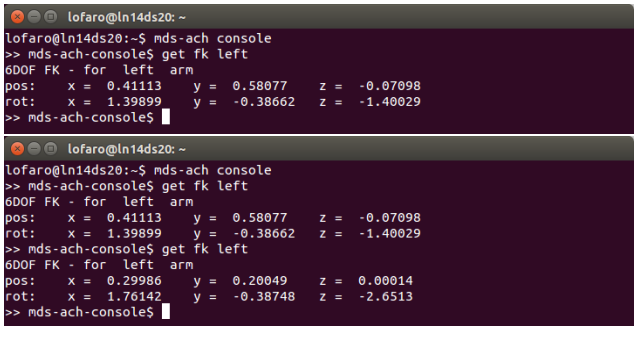


Figure 10: Expected MDS-Ach Console window for (>> mds-ach-console$ get fk left). (TOP) Before running. (BOTTOM) After running.

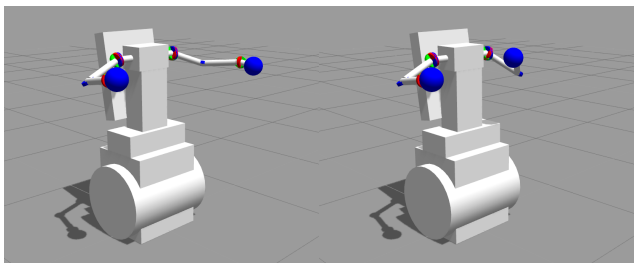


Figure 11: Expected simulator window for ($ python mds_ik_one.py). (LEFT) Before running. (RIGHT) After running.

8.1.9 Full Code

The full example can be found below as well in the file:

`mds-ach/examples/mds_ik_one.py`

8.2 Built-in IK Controller (C/C++)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system works by starting to solve the IK equations when you post a new desired position on the IK reference channel. Appendix G has a step by step of how you do this using C/C++ which is identical in methodology to the Python implimentaiton in Section 8.1.

8.3 Making a Box using Inverse Kinematics (Python)

This section shows you how to control the MDS robot via the built in IK module. This example shows us using solving a 3 DOF IK. The method used can be expanded to any DOF between 1 and 6.

The example given is the robot moving its left hand in a 0.5 m box 0.4 m away from the origin. The hand will not move in the x plane, only in the y and z.

8.3.1 Making the Box

The following points will be hit in order: $(0.4, 0.45, 0.25) \rightarrow (0.4, -0.05, 0.25) \rightarrow (0.4, -0.05, -0.25) \rightarrow (0.4, 0.45, -0.25)$ with all units in meters. Example Python code for state flow. Note state 0 $(0.3, 0.2, 0.0)$ is the initial state and will not be returned to.

Python:

```
if ii == 0:
  c = [0.3, 0.2, 0.0]
  ii = 1
elif ii == 1:
  c = [0.4, 0.45, 0.25]
  ii = ii+1
elif ii == 2:
  c = [0.4, -0.05, 0.25]
  ii = ii+1
elif ii == 3:
  c = [0.4, -0.05, -0.25]
  ii = ii+1
elif ii == 4:
  c = [0.4, 0.45, -0.25]
  ii = 1
```

8.3.2 Select Arm

The following selects the arm used for the IK. You may use "left" or "right" for the left and right arms respectively.

Python:

```
arm = 'left'
armi = -1
if arm == 'left':
  armi = mds.LEFT
if arm == 'right':
  armi = mds.RIGHT
```

8.3.3 Parse Desired Position

Parse the desired position for the end-effector to 6 DOF array.

Python:

```
dof = 3
eff = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
for i in range(0,dof):
  eff[i] = c[i]
```

8.3.4 Set Desired Position

Sets the desired position for the end-effector. ikc.arm[].ik_method is the number of DOF that you will be controlling. ikc.arm[].t_* and ikc.arm[].r_* is the desired position and rotation of the end effector respectively.

Python:

```
if armi >= 0:
  ikc.move = armi
  ikc.arm[armi].ik_method = dof
  ikc.arm[armi].t_x = eff[0]
  ikc.arm[armi].t_y = eff[1]
  ikc.arm[armi].t_z = eff[2]
  ikc.arm[armi].r_x = eff[3]
  ikc.arm[armi].r_y = eff[4]
  ikc.arm[armi].r_z = eff[5]
```

8.3.5 Send Desired Position to be Solved

The following sends the desired position and ik method to the IK controller to attempt a solution.

Python:

```
k.put(ikc)
```

8.3.6 Running the Code

To run the code do the following from within the *mds-ach/examples* directory

Bash:

```
$ python examples/mds_ik_box_example.py
```

The expected terminal output can be seen in Figure 12.

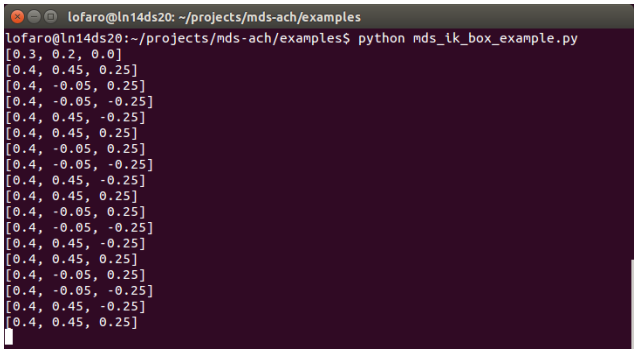


Figure 12: Expected window for ($ python examples/mds_ik_box_example.py). Time order is left to right, top to bottom.

The expected robot pose can be seen in Figure 13. The latter figure shows the virtual robot not the physical robot.

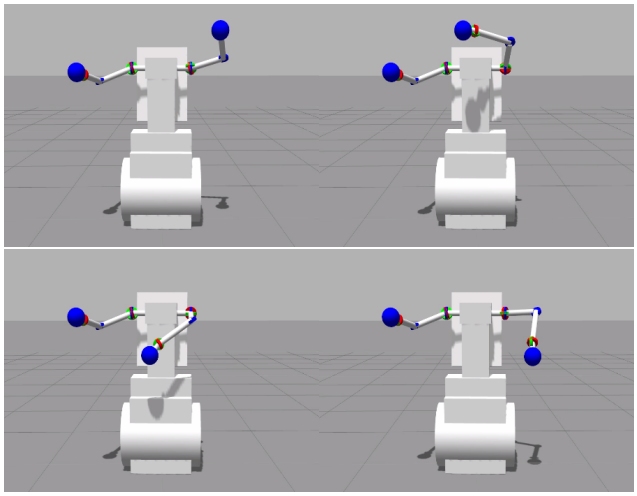


Figure 13: Expected robot poses for ($ python examples/mds_ik_box_example.py)

8.3.7 Full Code

The full example can be found in the file:

```
mds-ach/examples/mds_ik_box_example.py
```

8.4 Inverse Kinimatics (IK) API

The Inverse Kinematic (IK) API utilizes the Inverse Jacobian method to solve for the joint-space values given an end-effector position of each of the arms. The IK solver is capable of solving for any and all of the six degrees of freedom of the end-effector i.e. translation (x, y, x) and rotation $(\theta_x, \theta_y, \theta_z)$ while keeping the non-constrained position/rotations in Null space. The number of steps and error range can be specified by the user.

Appendix H gives an example of how to get and set the forward and inverse kinematics via the mds_ik API. The goal of this tutorial is to solve for the joint-space values given the desired work-space values as found in Table 3.

9 Conclusion

In conclusion we have made a middleware called MDS-Ach that enables the legacy MDS Robot to be used with modern day robot software, thus extending its life as a research robot. Low-latency non-head-of-line blocking FILO shared memory and network connectivity is used to share data between real-time processes. SSH tunneling is used if a secure network connection between controllers is required. Built-in collision avoidance, inverse kinematics, and support for multiple programming languages was implemented to expand usability to our non-hardware-focused partners. Finally, a ROS interface was developed with specific focus on making it ROS 2.0 compatible to enable the use of the extensive ROS ecosystem. These combined contributions allowed MDS-Ach to significantly extend the research life of the MDS Robot.

Acknowledgment This work was performed in part at the Naval Research Laboratory under the project Adaptive Real-Time Algorithms for Multiagent Cooperation in Adversarial Environments. The views, positions and conclusions expressed herein reflect only the authors opinions and expressly do not reflect those of the Naval Research Laboratory, nor those of the Office of Naval Research.

References

[1] M. Quigley, K. Conley, B. P. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, 2009.

[2] C. Breazeal, M. Siegel, M. Berlin, J. Gray, R. A. Grupen, P. Deegan, J. Weber, K. Narendran, and J. McBean, "Mobile, dexterous, social robots for mobile manipulation and human-robot interaction," in *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2008, Los Angeles, California, August 11-15, 2008, New Tech Demos Proceedings*. ACM, 2008, p. 27. [Online]. Available: http://doi.acm.org/10.1145/1401615.1401642

[3] Y. Maruyama, S. Kato, and T. Azumi, "Exploring the performance of ros2," in *Proceedings of the 13th International Conference on Embedded Software*, ser. EMSOFT '16. New York, NY, USA: ACM, 2016, pp. 5:1–5:10. [Online]. Available: http://doi.acm.org/10.1145/2968478.2968502

[4] D. Calisi, A. Censi, L. Iocchi, and D. Nardi, "OpenRDK: a framework for rapid and concurrent software prototyping," in *Proceedings of Int. Workshop on System and Concurrent Engineering for Space Applications (SECESA)*, Nov. 2008.

[5] T. H. J. Collett and B. A. Macdonald, "Player 2.0: Toward a practical robot programming framework," in *in Proc. of the Australasian Conference on Robotics and Automation (ACRA*, 2005.

[6] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: Yet another robot platform," *International Journal of Advanced Robotics Systems, special issue on Software Development and Integration in Robotics*, vol. 3, no. 1, 2006.

[7] C. Ct, Y. Brosseau, D. Ltourneau, C. Raevsky, and F. Michaud, "Robotic software integration using marie," *International Journal of Advanced Robotic Systems*, vol. 3, no. 1, pp. 055–060, March 2006. [Online]. Available: http://www.ars-journal.com/International-Journal-of-Advanced-Robotic-Systems/Volume-3/055-060.pdf

[8] N. Ando, T. Suehiro, K. Kitagaki, T. Kotoku, and W.-K. Yoon, "RT-Component Object Model in RT-Middleware Distributed Component Middleware for RT (Robot Technology)," pp. 457–462, Jun. 2005.

[9] A. Makarenko and A. Brooks, "Orca: Components for robotics," in *In 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.

[10] M. Henning, "A new approach to object-oriented middleware," *IEEE Internet Computing*, vol. 8, no. 1, pp. 66–75, Jan 2004.

[11] S. Rowe and C. R. Wagner, "An introduction to the joint architecture for unmanned systems (jaus)," *Ann Arbor*, vol. 1001, p. 48108.

[12] N. Dantam, D. Lofaro, A. Hereid, P. Oh, A. Ames, and M. Stilman, "The ach library: A new framework for real-time communication," *Robotics Automation Magazine, IEEE*, vol. 22, no. 1, pp. 76–85, March 2015.

[13] D. M. Lofaro, C. Ward, M. Bugajska, and D. Sofge, "Extending the life of legacy robots: Mds-ach, a real-time, process based, networked, secure middleware based on the x-ach methodology," in *15th International Workshop on Advanced Motion Control (IEEE-AMC2018)*, March 2018.

[14] A. Perez, M. Orsag, and D. Lofaro, "Design, implementation, and control of the underwater legged robot aquashoko for low-signature underwater exploration," in *2018 15th IEEE International Conference on Ubiquitous Robots (UR)*, 2018.

[15] J. Brown and B. Martin, "How fast is fast enough? Choosing between Xenomai and Linux for real-time applications," in *Twelfth Real-Time Linux Workshop on October 25 to 27, in Nairobi, Kenya*, 2010.

[16] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, Sep 2004, pp. 2149–2154.

[17] R. Smith, "Open dynamics engine," 2008. [Online]. Available: http://www.ode.org/

[18] S. R. Buss, "Introduction to inverse kinematics with jacobian transpose, pseudoinverse and damped least squares methods."

[19] D. Lofaro, R. Ellenberg, P. Oh, and J. Oh, "Humanoid throwing: Design of collision-free trajectories with sparse reachable maps," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, Oct 2012, pp. 1519–1524.

[20] D. M. Lofaro, "Secure robotics," in *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Aug 2016, pp. 311–313.

[21] D. Lofaro, "Mds-ach," in *https://github.com/thedancomplex/mds-ach*, 2015.

10 APPENDIX

A Daemon Console

This section shows the specific console inputs available for the MDS-Ach system.
Console Input: $ mds-ach - Shows command options
Console Input: $ mds-ach start - Start all channels and processes and console

- (no-arg) : Starts MDS-Ach system on CAN Bus 0 "CAN0"
- nocan : Starts MDS-Ach system with no output to real CAN Bus. Will output to the virtual can "VCAN42" instead

Console Input: $ mds-ach console - Starts the human interface console for MDS-Ach
Console Input: $ mds-ach stop - Close all channels and processes
Console Input: $ mds-ach make - makes all the MDS channels
Console Input: $ mds-ach kill - Emergency kill the daemon process
Console Input: $ mds-ach killall - Emergency kill the daemon process and removes all ACH channels
Console Input: $ mds-ach resetbus - Resets the Bus
Console Input: $ mds-ach remote - Starts a remote connection to xxx.xxx.xxx.xxx via achd
Console Input: $ mds-ach sim - Starts the sim in gazebo

- (no-arg) : Starts the sim in gazebo
- kill : Kills gazebo sim

Console Input: $ mds-ach changerobot - Changes the robot's configuration file/anatomy

- (no-arg) : No Change
- isaac : Changes to Isaac's anatomy
- lucas : Changes to Lucas' anatomy
- octavia : Changes to Octavia's anatomy

B Startup Procedure

To start the MDS-Ach daemon select one of the two options. Note: Both Option 1 and Option 2 will start the following processes:

- mds-daemon : *primary control for the MDS robot*
- mds-filter : *filtering process to allow for step inputs*
- mds_ik_module (python) : *inverse kinematics (ik) controller for the MDS*

B.1 (Option 1) With physical MDS robot:

Bash:

```
$ mds-ach start
```

B.2 (Option 2) With virtual MDS robot:

Bash (Start with no CAN Bus):

```
$ mds-ach start nocan
```

Bash (Start Simulator):

```
$ mds-ach sim
```

The simulator can be run with either (Option 1) or (Option 2) above.

C Examples

The following are examples of how to do basic operations using the MDS-Ach system on both the real robot and the simulator.

Run MDS-Ach Daemon:

Bash:

```
$ mds-ach start
```

The resulting terminal windows should look like Figure 14.

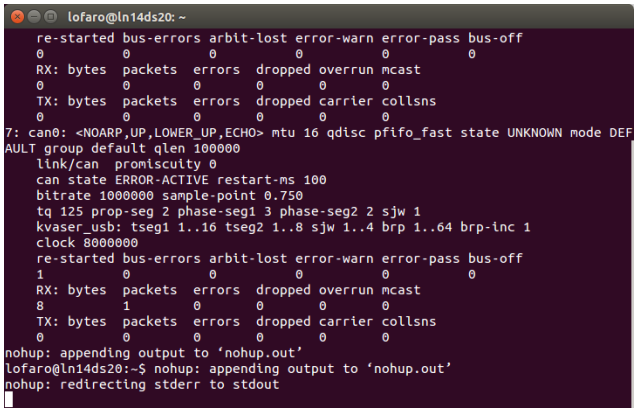


Figure 14: Expected window for ($ mds-ach start)

Run MDS-Ach Daemon with simulator only:

Bash:

```
$ mds-ach start nocan
```

This command will start the MDS-Ach daemon. This should be run on the computer where the simulator is located. The MDS-Ach daemon (mds-daemon) will run in the background even if the terminal session is closed. This will run with the simulator and not with the real robot. Use this mode if you only want to use the simulator.

The resulting terminal windows should look like Figure 15.

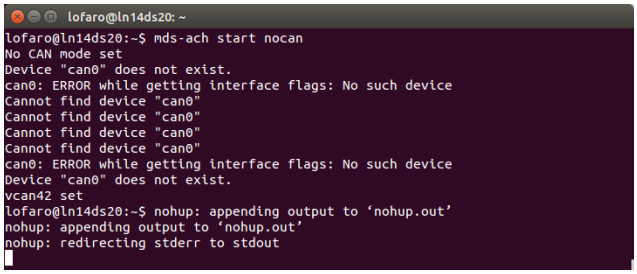


Figure 15: Expected window for ($ mds-ach start nocan)

Run MDS-Ach simulator:

Bash:

```
$ mds-ach sim
```

Once the MDS-Ach Daemon is started (see above) you can run the simulator. The simulator is open loop in the respect that it does not feed back information to the MDS-Ach system. It is used as a visual representation of the robot for debugging and initial controller testing.

The resulting terminal windows should look like Figure 16.



Figure 16: Expected window for ($ mds-ach sim)

The resulting simulator windows should look like Figure 6.

D MDS-Ach Console

The MDS-Console utility allows the user to get and set joint space values via the command line. It also allows the user to get the work-space end-effector position (6 DOF) and set the work-space end-effector position (3, 4, 5, and 6 DOF).

Prerequisites:

The MDS-Ach system must be running prior to running the MDS-Console

Startup:

Bash:

```
$ mds-ach console
```

To start the MDS-console run the above command. The expected terminal can be seen in Figure 17.

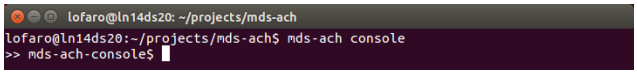


Figure 17: Expected window for ($ mds-ach console)

Commands:

This section shows the commands available to the MDS-Console. Note: when in MDS-Console the console will shows the following in the terminal:

Bash/MDS-Ach Console:

```
>> mds-ach-console$
```

Function: goto:

The goto command will tell the joint (in joint space) what position (in radians) where to go. The usable joints and abbreviations can be found in Table 2.

Bash/MDS-Ach Console:

```
>> mds-ach-console$ goto <joint> <value>
```

This will set the Right Shoulder Pitch (RSP / RightShoulderExt) to a value of -0.123 *rad*. The expected terminal can be seen in Figure 18.

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ goto RSP -0.123
```

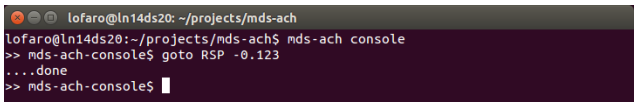


Figure 18: Expected window for (>> mds-ach-console$ goto < *joint* > < *value* >) using example (>> mds-ach-console$ goto RSP -0.123)

The "get" command gets the joint space position of the < *joint* > in *radians*. The usable joints and abbreviations can be found in Table 2.

Function: get:

Bash/MDS-Ach Console:

```
>> mds-ach-console$ get <joint>
```

This will get the reference and state of the Torso yaw (WST / TorsoPan). The expected terminal can be seen in Figure 19.

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ get WST
```

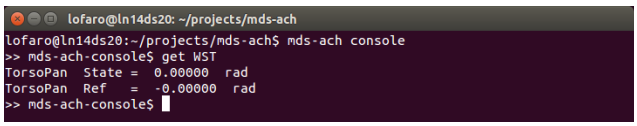


Figure 19: Expected window for (>> mds-ach-console$ get < *joint* >) using example (>> mds-ach-console$ get WST)

Function: get fk:

The "get fk" command gets the work-space position of the left or right end-effector in 6 DOF coordinates (meters and radians) with the origin being the intersection of the robot's neck and shoulder. The < *arm* > options are: **left** and **right** for the left and right arm respectively.

Bash/MDS-Ach Console:

```
>> mds-ach-console$ get fk <arm>
```

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ get fk left
```

This will get the 6 DOF work space position of the left end-effector. The expected terminal can be seen in Figure 20.

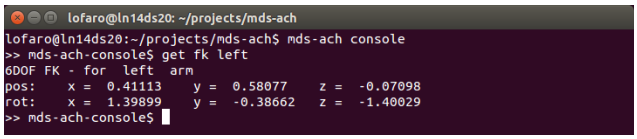


Figure 20: Expected window for (>> mds-ach-console$ get fk < *arm* >) using example (>> mds-ach-console$ get fk left) where (pos) is the position in meters and (rot) is the rotation about x,y,z in radians.

Function: ik:

Bash/MDS-Ach Console:

```
>> mds-ach-console$ ik <arm> <dof> <param 1>
                                  ... <param N>
```

The "ik" command utilizes the inverse kinematic controller (mds_ik_module) to solve 1-6 DOF inverse kinematic solutions for the MDS robot's left or right end-effector. This module uses the Inverse Jacobian Inverse Kinematic solver method. If the desired workspace location is too far away such that the IK controller cannot reach it in 1000 iterations, it will discontinue attempting to find a solution and return without a reply. You may then try another point closer to that of the current end-effector point.

< **arm** > : has the options of "left" and "right." and denote the left and right end-effector respectively.

< **dof** > : denotes the number of degrees of freedom you will be controlling using the inverse kinematic controller.

< **param 1** > ... < **param N** > : denotes the positions and orientations for the < *arm* >. The number of parameters must equal that of < *dof* >. The order must be as follows (Table 5):

Table 5: Inverse Kinematic Parameter Order

Param #	Definition	Abbreviation
1	Position in X	p_x
2	Position in Y	p_y
3	Position in Z	p_z
4	Rotation in X	θ_x
5	Rotation in Y	θ_y
6	Rotation in Z	θ_z

Bash/MDS-Ach Console (Example):

```
>> mds-ach-console$ ik left 3 0.3 0.2 0.0
```

This will find a joint space solution using IK methods for the desired end-effector position to be (0.3 m, 0.2 m, 0.0 m) in (x, y, z). The angle about all axes is in the null space.

The expected output (with running get fk left before and after to show the effect) can be seen in Figure 21.

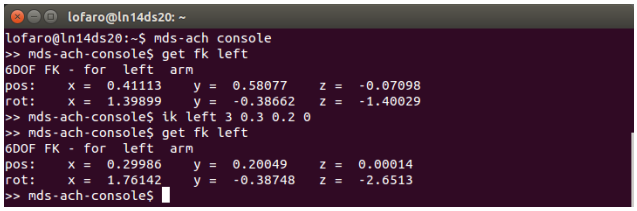


Figure 21: Expected window for (>> mds-ach-console$ ik < *arm* > < *dof* > < *param* 1 > ... < *param N* >) using example (>> mds-ach-console$ ik left 3 0.3 0.2 0.0)

E Software Interface

The MDS-Ach system currently works with the C/C++ and Python programming languages. This section describes the required libraries for each of the latter languages.

Python imports:

The following are required imports for the MDS-Ach system while using python: mds_ach, ach, and mds. Other imports are for the given controller implementation.

Python:

```
#!/usr/bin/env python
import mds_ach as mds
import ach
```

C/C++ includes:

The following are required imports for the MDS-Ach system while using python: mds_ach, ach, and mds. Other imports are for the given controller implementation.

C/C++:

```
#include <mds.h>
#include <ach.h>
```

Required library:

MakeFile:

```
-lach
```

C/C++ MakeFile example with required libraries: The following is an example make file for a C implementation of a controller. Please note it utilizes the required **-lach** library.

MakeFile:

```
default: all
CFLAGS := -I./include -g -std=gnu99
CC := gcc
BINARIES := mds-simple-demo
LIBS := -lach -lrt -lm

all : $(BINARIES)

mds-simple-demo: src/mds-simple-demo.o
	$(CC) -o $@ $< $(LIBS)

%.o: %.c
	$(CC) $(CFLAGS) -o $@ -c $<

clean:
	rm -f $(BINARIES) src/*.o
```

F Operating in Joint-Space

This section shows you how to setup the Ach channels for communication with the MDS-Ach system and how to control the robot in joint-space while using the smoothing filter process.

F.1 Control one joint/DOF (Python)

This section shows you how to setup the Ach channels for communication with the MDS-Ach system and how to control one joint of the robot in joint-space while using the smoothing filter process. Specifically we will set the joint-space values as those seen in Table 6. Note that to run this example you will need the libraries for MDS-Ach as seen in Section E.

Table 6: Set the joint space values for the following joint

Name	Alternate Name	Value (rad)
LSP	LeftShoulderExt	-0.123

Open Ach Channels:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that ach.Channel() takes a string as an input. Here we opened two channels:

- s : state channel
- r : reference channel

Python:

```
s = ach.Channel(mds.MDS_CHAN_STATE_NAME)
r = ach.Channel(mds.MDS_CHAN_REF_NAME)
```

Create Required Data Structures:

C-Type data structures are used to pass data between our controllers. Below we create three well defined structures for the state and the reference channels. These structures are defined in mds_ach.py and mds.h which is located in your python and include paths.

- state : state channel of type MDS_STATE
- ref : reference channel of type MDS_REF

Python:

```
state = mds.MDS_STATE()
ref = mds.MDS_REF()
```

Get Joint ID:

To command a joint you must get the ID of the joint. The ID numbers are defined in the anatomy.xml configuration file. You can use the joint abbreviations in Table 18 to find the ID number.

Python:

```
# Get address of LSP
jntn = mds.getAddress('LSP',state)
print 'Address = ', jntn
```

Queue New Motor Position:

You can set a new desired motor angle by setting the reference channel. In this case we are using the filter channel which is the safest one to use due to the velocity and acceleration limiting. Please note that this does NOT send the command to the motor. It queues the values for them to be sent to the motors. They are only sent to the motors after they are "put" on the ach channel.

In the example below we are setting the 'LSP' by using the ID number from above to a joint-space angle of -0.123 *rad*.

Python:

```
# Set LSP reference to -0.123 using the
# filter controller
ref.joint[jntn].ref = -0.123
```

Set Motor Position (put):

Once all joints are set in the structure you can "put" it on the proper ach channel. Please note that even if you did not set a motor value it will still be put on the channel along with the rest of the data structure. It is best practice to read the latest channel, then modify what you want to change, then put the modified structure on the channel. This will help with not sending the robot to unintended configurations.

Python:

```
# Command motors to desired references
# (post to ach channel)
r.put(ref)
```

Close Ach Channels on Exit:

Though not required it is good practice to close your unused ach channels upon exit of your controller.

Python:

```
# Close channels
s.close()
r.close()
```

Running the Code: To run the code do the following from within the *mds-ach/examples* directory

Bash:

```
$ python mds_simple_demo_python_1_DOF.py
```

The expected terminal out can be seen in Figure 22.

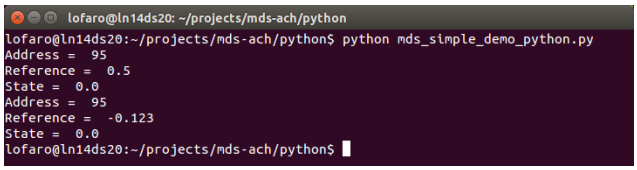


Figure 22: Expected window for the python 1-DOF example including additional print statements

You can also monitor the change using the MDS-Ach Read utility and/or the simulator/real robot.

Full Code:

The full example can be found below as well in the file:

`mds-ach/examples/mds_simple_demo_python_1_DOF.py`

F.2 Control Two joints/DOF (C/C++)

This section you will set 2 separate DOFs to two different values and see the results. Specifically we will set the joint-space values as those seen in Table 7. Note that to run this example you will need the libraries for MDS-Ach as seen in Section E. We will set the following:

Table 7: Set the joint space values for the following joint

Name	Alternate Name	Value (rad)
LEP	LeftElbowFlex	-0.2
RSP	RightShoulderExt	0.1

Open Ach Channels:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that ach.Channel() takes a string as an input. Here we opened three channels.

- chan_state : state channel
- chan_ref : reference channel

C/C++:

```
int r = ach_open(&chan_ref,
                 MDS_CHAN_REF_FILTER_NAME, NULL);
assert( ACH_OK == r);

r = ach_open(&chan_state,
                 MDS_CHAN_STATE_NAME, NULL);
assert( ACH_OK == r);
```

Create Required Data Structures:

C-Type data structures are used to pass data between our controllers. Below we create three well defined structures for the state and the reference channels. These structures are defined in mds_ach.py and mds.h which is located in your python and include paths.

- state : state channel of type MDS_STATE
- ref : reference channel of type MDS_REF

C/C++:

```
mds_ref_t H_ref;
mds_state_t H_state;
memset( &H_ref,   0, sizeof(H_ref));
memset( &H_state, 0, sizeof(H_state));
```

Get Up-To-Date Reference:

It is important that you set your initial reference structure with the current values of the reference channel. This is because when you do command the joints you command them all at once, even if you did not change their value. See below for how to get the latest reference values.

C/C++:

```
int r = ach_get( &chan_ref, &H_ref,
                 sizeof(H_ref), &fs,
                 NULL, ACH_O_LAST );
    if(ACH_OK != r) {
        if(debug) {
          fprintf(stderr, "Ref r = %s\n",
                  ach_result_to_string(r));}}
    else{assert( sizeof(H_ref) == fs);}
```

Get Joint ID:

For the moment when using C/C++ the joint IDs are hard coded in the mds.h include file. You may use normal humanoid acronyms or NRL legacy acronyms.

Queue New Motor Positions:

You can set a new desired motor angle by setting the reference channel. In this case we are using the filter channel which is the safest one to use due to the velocity and acceleration limiting. Please note that this does NOT send the command to the motor. It queues the values for them to be sent to the motors. They are only sent to the motors after they are "put" on the ach channel.

In the example below we are setting the joints as defined in Table 7.

C/C++:

```
H_ref.joint[LEP].ref = -0.2;
H_ref.joint[RSP].ref = 0.1;
```

Set Motor Position (put):

Once all joints are set in the structure you can "put" it on the proper ach channel. Please note that even if you did not set a motor value it will still be put on the channel along with the rest of the data structure. It is best practice to read the latest channel, then modify what you want to change, then put the modified structure on the channel. This will help with not sending the robot to unintended configurations.

C/C++:

```
/* Write to the feed-forward channel */
ach_put( &chan_ref, &H_ref, sizeof(H_ref));
```

Running the example:

To run the example compile then run the resulting executable mds-simple-demo.

Bash:

```
$ ./mds-simple-demo
```

The resulting output can be seen using the MDS-Ach Read utility (Figure 7) and/or the simulator/real robot (Figure 8).

Full Code:

The full example can be found in the file:

`mds-simple-demo/src/mds-simple-demo.c`

G Built-in IK Controller (C/C++)

The MDS-Ach system has a built-in inverse kinematics (IK) controller for both the right and left arms. This controller automatically runs when the MDS-Ach system is started. The controller solves for one arm at a time.

The system works by starting to solve the IK equations when you post a new desired position on the IK reference channel. Below is a step by step of how you do this using C/C++.

C/C++ Includes:

The following are required imports for the MDS-Ach system while using python: mds.h and ach.h. Other imports are for the given controller implementation.

C/C++:

```
// for mds
#include <mds.h>
// for ach
#include <ach.h>
```

Create Open Ach Channel for IK:

Ach channels are how you communicate with MDS-Ach. You simply write data to the channel and the robot can read the data in newest to oldest order. The section below shows you how to open an Ach channel. Please note that ach_open() takes a string as an input. Here we opened one channel. This is a different channel then found in previous sections because it is only for the IK controller.

C/C++:

```
// open ik chan
int r = ach_open(&chan_ik,
                 MDS_CHAN_IK_NAME, NULL);
assert( ACH_OK == r);
```

Make IK Structure:

Similar to controlling the robot in joint-space you need to set a reference structure to the desired work-space position. For this you need to initialize the structure. See below for the initialization of the work-space structure.

C/C++:

```
// Make new IK structure
mds_ik_t H_ik;
memset( &H_ik,    0, sizeof(H_ik));
```

Setting the DOF controlled:

When using the built-in IK controller you need to set the number of DOF that you are controlling. With this controller you are required to set the DOF in the following order: $p_x, p_y, p_z, \theta_x, \theta_y, \theta_z$ where p_n is the position on axis n and θ_n is the rotation about axis n. For example if you DOF is set to 4 you are controlling p_x, p_y, p_z, and θ_x. If you are controlling 2 you will control p_x and p_y. In the example below we are controlling 3, i.e. p_x, p_y, and p_z. This order can also be seen in Table 5.

C/C++:

```
// Set the amount of DOF you want to control
dof = 3
```

We set these to the values in Table 3.

θ_x, θ_y, and θ_z are in the Null Space because we do not care where they are as long as the first three parameters are met. We can set these values to what ever we want and they will be ignored. In this case we set them to zero.

C/C++:

```
/* Set values for work-space in
   [x, y, z, rx, ry, rz] order */
double eff[6] = {0.3, 0.2, 0.0, 0.0, 0.0, 0.0};
```

Choosing Arm for IK:

Here we pick the arm for the IK. Our options are set as enums in the mds.h include. All options for right and left arms can be found in Table 8.

Table 8: Definitions for left and right arms using mds_ach.py in Python

Arm	C/C++ Definition
Left	LEFT
Right	RIGHT

Here we set the arm to the left arm.

C/C++:

```
// set arm
int arm = LEFT;
```

Set IK Structure:

Just as in the joint-space method, we need to the values in our structure before we send it to the robot. Here we set all of the parameters from above to the structure ikc that we created.

C/C++:

```
// Put setting into ik structure
H_ik.move = arm;
H_ik.arm[arm].ik_method = dof;
H_ik.arm[arm].t_x = eff[0];
H_ik.arm[arm].t_y = eff[1];
H_ik.arm[arm].t_z = eff[2];
H_ik.arm[arm].r_x = eff[3];
H_ik.arm[arm].r_y = eff[4];
H_ik.arm[arm].r_z = eff[5];
```

Command the Robot:

Just as with the joint-space controller you need to "put" the data on the ACH channel. Unlike the joint-space controller it will not move as soon as you send it. The controller will first have to solve the IK. Upon finding the solution the controller will send it to the robot. If a solution is found it will take anywhere between 0.1 and 5.0 seconds. If there is no solution found the robot will not move. Note: the robot limits its self to 1000 search iterations for an IK solution.

C/C++:

```
// put on to ACH channel
ach_put( &chan_ik, &H_ik, sizeof(H_ik));
```

Running the Code:

To run the code do the following within the **mds-simple-demo-ik** folder.

Bash:

```
$ make clean
$ make
$ ./mds-simple-demo-ik
```

To check that the IK worked you can run the FK in the MDS-Ach console and/or run the simulator. The before and after of the MDS-Ach console is found in Figure 23 and Figure 24 respectively.

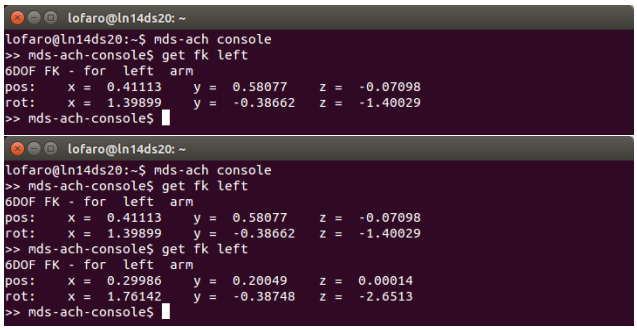


Figure 23: Expected MDS-Ach Console window for (>> mds-ach-console$ get fk left). (TOP) Before running. (BOTTOM) After running.

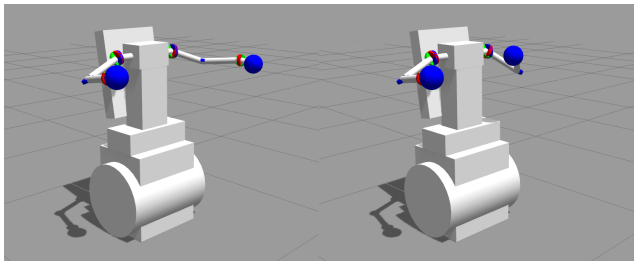


Figure 24: Expected simulator window for ($ ./mds-simple-demo-ik) and/or ($ python mds_ik_solver_example.py). (LEFT) Before running. (RIGHT) After running.

Full Code:

The full example can be found in the file:

`mds-simple-demo/src/mds-simple-demo-ik.c`

H Inverse Kinematics API

This section will show you an example of how to get and set the forward and inverse kinematics via the mds_ik API. The goal of this tutorial is to solve for the joint-space values given the desired work-space values

as found in Table 3. This API is written for use with Python but also has hooks for C/C++.

Required Imports:

Below are the required imports for using the forward/inverse kinematics portion of the mds API. These will be located in the python 2.7 path.

Python:

```
import mds_ach as mds
import ach
import mds_ik as ik
import mds_ik_include as ike
from mds_ach import *
```

Open Ach Channels:

You will need the state and the reference (filtered) channels to solve and set the IK using the API. The state is used to get the current location of the robot. The reference is used to set the joint-space values once they are found.

Python:

```
# Open Ach Channels
r = ach.Channel(mds.MDS_CHAN_REF_FILTER_NAME)
s = ach.Channel(mds.MDS_CHAN_STATE_NAME)
```

Make Structures:

Make the structures for the state and the reference channels. This will hold your latest state data and the latest set reference to the reference filter channel.

Python:

```
# Make Structs
state = mds.MDS_STATE()
ref = mds.MDS_REF()
```

Get Latest Data:

Get the latest data from the state and reference channels.

Python:

```
# Get the latest on the channels
[status, framesize] = s.get(state, wait=False,
                            last=True)
[status, framesize] = r.get(ref, wait=False,
                            last=True)
```

Set Desired Work-Space Position:

Set the desired work-space position as found in Table 3. We set all 6 DOF here however only the first three are required due to our desired work-space position. Later sections removes the extra DOFs.

Python:

```
# Desired position
eff_end = np.array([0.3,0.2,0.0,0.0,0.0,0.0])
```

Set Desired Arm:

The ik.getIK() function takes a string to define the left or right arm. Thus we set our "arm" as a string. The options for the latter can be found in Table 9.

Table 9: Definitions for left and right arms using mds_ik.py in Python

Arm	MDS IK (Python) Definition
Left	'left'
Right	'right'

Python:

```
# Define Arm
arm = 'left'
```

Get Current Joint-Space Pose:

Next we have to get the current joint-space pose of the robot and putting it into a 6 DOF array. This is done by looking for the index of each joint and then putting them in an array. The array must be in the order defined in Table 10.

Python:

```
# Get current joint space pose of arm
jntn = mds.getAddress('LSP',state)
j0 = state.joint[jntn].ref
jntn = mds.getAddress('LSR',state)
j1 = state.joint[jntn].ref
jntn = mds.getAddress('LSY',state)
j2 = state.joint[jntn].ref
jntn = mds.getAddress('LEP',state)
j3 = state.joint[jntn].ref
jntn = mds.getAddress('LWY',state)
j4 = state.joint[jntn].ref
jntn = mds.getAddress('LWR',state)
j5 = state.joint[jntn].ref
eff_joint_space_current=[j0,j1,j2,j3,j4,j5]
```

Table 10: Arm joint-space pose order.

Array Index	Joint Definition	Short Name (left)	Short Name (right)
0	Shoulder Pitch	LSP	RSP
1	Shoulder Roll	LSR	RSR
2	Shoulder Yaw	LSY	RSY
3	Elbow Pitch	LEP	REP
4	Wrist Yaw	LWY	RWY
5	Wrist Roll	LWR	RWR

Set DOF and Order of Desired Work-Space Position:

We now set the DOF of the desired work-space (acceptable numbers are 1-6) and the order in which we

input them. Using the MDS IK API we can input the desired work-space coordinates in any order we desire. The order is defined by an array of string that we call 'order'. The valid inputs for this array can be found in Table 11.

Table 11: Valid Inputs for "order" Array

Definition	Input to Array (String)
Position in x (p_x) (meters)	'p_x'
Position in y (p_y) (meters)	'p_y'
Position in z (p_z) (meters)	'p_z'
Rotation about x (θ_x) (radians)	't_x'
Rotation about y (θ_y) (radians)	't_y'
Rotation about z (θ_z) (radians)	't_z'

Here we have 3 DOF with an order of position in x, y, then z. Note: This order MUST match that of the order in your desired work-space position array in Section H.

Python:

```
# set the dof and the order (dof = 3)
dof = 3
order = ['p_x','p_y','p_z']
```

Set Error:

Here we set the error for the IK solver. The error is a 3D array with the atrobutes found in Table 12.

Table 12: Error Array Index Definitions

Index	Definition
0	Max change in angular position (θ) per iteration (rad)
1	Max change in linear position (xyz) per iteration (m)
2	Max linear Error ($m^2 + rad^2$)

The max change in angular position is the maximum distance in radians that the solver will move per iteration. The max change in translational position is the maximum distance translated in x, y, z per iteration in meters. The max linear error (e_{max}) is the linear distance in N DOF that disregards units. The error in the end-effectors actual position vs its desired position (e_{eff}) can be found via the equation below.

$$e_{eff} = \left(\sum_{i=0}^{N-1} (d_{des}[i] - d_{act}[i])^2 \right)^{\frac{1}{2}} \quad (5)$$

where d_{des} is the desired pose and d_{act} is the actual pose. The system will say the IK has been solved if $e_{eff} \leq e_{max}$.

To set the desired errors make the 3D array as seen below. If no error is input the default values of $[0.01, 0.01, 0.01]$ will be used.

Python:

```
# set the allowable error
err = np.array([0.01,0.01, 0.01])
```

Set Max Solving Iterations:

Due to the nature of the IK method utilized the system can fall into local minimum or a hysteresis. This will cause the system to attempt to solve indefinitely. To avoid this use add a maximum number of iterations to try when solving the IK. This number is set below. If not set it will use the default value of 1000.

Python:

```
# Set solving step number max
stepNum = 1000
```

Limit input to desired DOF:

Here you limit the size of the desired end-effector position vector to the 'dof' that we defined in Section H.

Python:

```
# Deisred position for only the dof we want
eff_end = eff_end[:dof]
```

Solving the IK:

Now we can solve the IK via the use of ik.getIK() in the mds_ik python module.

Python:

```
# Solve IK
jnt_return = ik.getIK(eff_joint_space_current,
                      eff_end, order, arm, err,
                      stepNum)
```

The output of this is a 2D array where index 1 returns a -1 if there is no IK solved. Index 0 is the array of the end-effectors joint-space values for the given desired work-space position. The order of the joint-space values are the same as what is found in Table 10.

Python:

```
# returns in the following order
# Joint space return =
# [LSP, LSR, LRY, LEP, LWY, LWR]
eff_joint_space_current = jnt_return[0]
```

Putting the IK solution on the robot:

To put the IK solution on the robot we map the received joint space values on the reference (ref) structure of the robot's MDS-Ach system. Again the order of the joint-space values are the same as what is found in Table 10. In this case we are setting the left arm.

Python:

```
# Set to joint space
jntn = mds.getAddress('LSP',state)
ref.joint[jntn].ref=eff_joint_space_current[0]
jnt n= mds.getAddress('LSR',state)
ref.joint[jntn].ref=eff_joint_space_current[1]
jntn = mds.getAddress('LSY',state)
ref.joint[jntn].ref=eff_joint_space_current[2]
jnt n= mds.getAddress('LEP',state)
ref.joint[jntn].ref=eff_joint_space_current[3]
jntn = mds.getAddress('LWY',state)
ref.joint[jntn].ref=eff_joint_space_current[4]
jntn = mds.getAddress('LWR',state)
ref.joint[jntn].ref=eff_joint_space_current[5]
```

Commanding the robot:

Just as in previous sections, we have to "put" the resulting structure on the reference (filtered) channel before the robot will move. Please note that we are using the filtered channel the the robot will not "jerk" during the joint-space step input operation.

Python:

```
# Send to the robot
r.put(ref)
```

Getting the Forward Kinematics from joint-space pose:

We can get the work-space position of the arm via the forward kinematics (FK) by utilizing the joint-space pose the the ik.getFkArm() function of the mds_ik module.

Python:

```
# get FK of arm
A = ik.getFkArm(eff_joint_space_current,arm)
eff_end_ret=ik.getPosCurrentFromOrder(A,order)
```

ik.getFkArm() will return a 4x4 matrix which includes the rotation and translations components. The ik.getPosCurrentFromOrder() returns the work-space position and orientation in the same order as requested by "order". This will return a 1xN array where N is equal to the length of your "order". You can find the error by doing the following.

Python:

```
# find different in desired vs actuial pos
eff_end_dif = eff_end - eff_end_ret
```

Running the code:

To run the code enter the "examples" directory of the mds-ach project and run the following command:

Bash:

```
$ python mds_ik_solver_example.py
```

The resulting console output should look like Figure 25.

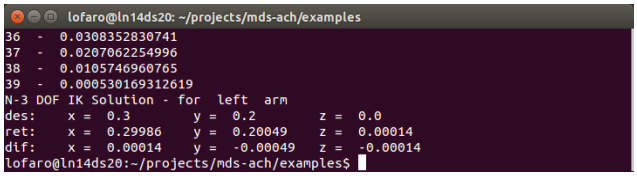


Figure 25: Expected window for ($ python mds_ik_solver_example.py)

The resulting simulator output should look like Figure 24.

Full Code:

The full example can be found in the file:

`mds-ach/examples/mds_ik_solver_example.py`

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 52-63 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Using of the Flipped Classroom Learning with a Workshop Activity in Object-oriented Analysis and Design Course

Pakawan Pugsee*

Innovative Network and Software Engineering Technology Laboratory, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, 10330, Thailand

ARTICLE INFO

Article history:
Received: 09 July, 2018
Accepted: 09 September, 2018
Online: 18 September, 2018

Keywords:
Flipped classroom learning
Learner behavior
Learner opinions

ABSTRACT

A lot of research reported that the active learning, especially the flipped classroom learning has higher performance than the lecture-based learning called the traditional learning. Therefore, this research is experimented on the object-oriented analysis and design course. The objectives are to try the flipped classroom learning with the workshop activity in order to analyze learners' behavior and to evaluate the effectiveness of the learning system. The aim of the research is to determine the suitability of the flipped classroom learning on the object-oriented analysis and design course, and to find the advantages and limitations of such learning method. The scope of this research is to test on the object-oriented analysis and design course, Faculty of Science, Chulalongkorn University. CourseVille developed by Chulalongkorn University is used as the learning management system. Moreover, the learning materials are Thai lecture videos, and English documents, such as English PowerPoint slides. Furthermore, the examples of activities in the classroom are Thai discussion and group activities, including the learner presentation. The research mythology is to collect data from learners composed of the behavior of the learners, learners' test scores, the result of the workshop activity and the output of term project. In addition, the collected data from the questionnaires about learner opinions with the classroom observation result are used for analyzing the learner attitude of the flipped classroom learning, and the learners' scores in the flipped classroom learning are compared to those in the traditional learning. In conclusion, the research results show that the 90 percentage of learners agree that the flipped classroom learning is suitable for the object-oriented analysis and design course, and it has higher performance that the traditional lecture-based learning system.

1. Introduction

At present, the learner-centered teaching method combined with the active learning is accepted widely that mainly focuses on learners, and can increase the effectiveness of learning and teaching system. One famous active learning method is the flipped classroom learning focused on the learners as the center and this system changes the traditional learning system like the lecture-based teaching from listening lecture contents in the classroom and doing homework or exercises after in-class time to the learners can study by themselves from the knowledge resource like out-of-school learning, such as teaching videos, online learning materials. On the other hand, the learners will practice in-class time to study, and the instructors will provide assignments, problem questions or case studies for learners to have a brainstorm and exchange ideas for finding those answers under the guidance of the instructors in the classroom. In addition, the learners and the instructors can join learning activities in the classroom to improve their knowledge, abilities, and skills.

In the previous academic year, a traditional object-oriented analysis and design (OOAD) course was the lecture-based learning which instructors will lecture the lessons in the classroom. After that, the class spends the rest of time for the questions of the case study to practice object-oriented analysis and design skill. Therefore, there is too less time to discuss and share opinions among learners and instructors in the classroom for

*Pakawan Pugsee, Bangkok, Thailand, +66 2218 5170 & pakawan.p@chula.ac.th

implementation. Moreover, learners cannot make the decision or conclude the different answers for problems or questions about study contents, when learners do group work outside the class time. A few reasons are that some learners cannot apply the course content to real world problems, or they are not able to compromise their different ideas generated from the understanding content. Moreover, they cannot achieve the best solutions for the view of object-oriented analysis and design.

For these limitations of traditional OOAD course, the instructors should try the flipped classroom learning in this course to analyze results of using this new learning system for OOAD course. In addition, the workshop activity outside the classroom were included to practice the object-oriented analysis and design skill. The learners' behavior of preparing for class and evaluation scores was analyzed to find the results of using the flipped classroom learning with a workshop activity in OOAD course. Learners' opinion was also investigated for concluding that the flipped classroom learning system can encourage learners to learn and practice analysis and design skill more or not.

This research paper is an extension of work originally presented in conference [1]. Proposed research details will be described in the following sections. The flipped classroom learning system and research about it are explained in Section 2. The research methodology is demonstrated in Section 3. Next, the research results with some discussions and conclusions are described in Section 4 and Section 5, respectively. Then, the final section is the suggestions.

2. The Flipped Classroom Learning System

The flipped classroom [2-3] is a learning combination between online studying and teaching in the classroom. This learning system is different from the traditional lecture-based learning that the learners do some practices (e.g. class discussions, presentations or assignments) in class, after they listened to lectures before in-class time outside the classroom. So, the learners can study the courses' subjects by learning the online content lecture videos or other online media resources outside the classroom, and then they will do in-class activities (e.g. having discussions, finishing exercises) in the classroom to more understand the subject. It is also helpful for learners to develop their knowledge and skills to compromise on the interests the individual learners.

Moreover, online social community has currently be contained a lot of knowledge and useful information for learning. This leads to instructors and learners to have valuable resources to teach and study. Instructors have more powers to generate learning experiences for learners. Learners will also understand the subject concept deeply from these experiences, thus the knowledge is not restricted to that has been transferred from the instructors [4]. Furthermore, there is some research about the development of Hybrid learning systems like integrated learning and teaching in the classroom [5]. Online tools can assist learners to do assignments, for examples, finding the answers or the guidance for problem solving in class. In addition, discussions on topics related class lessons or case studies can be run as online groups and learners can search information from other resources for generating knowledge besides course materials.

Therefore, the flipped classroom [6-9] is new learning and teaching system that can improve the academic exam scores of learners and encourages learners to exchange ideas and discuss and about different course contents. In addition, the flipped classroom learning is flexible enough to integrate with other methods for achieving more effectiveness of learning and adapt to the class topics as much as possible. Nevertheless, there is the limitation of individual learners' time that they must spend to study the preparation before in-class time attending. For example, the research [10] applied the flipped classroom learning in computer programming course that learners had to study the definition and syntax of programming languages before class and get their hands-on practices for deeper learning taking place in the classrooms. It was found that the flipped classroom allows for more interaction during the face-to-face meetings, and for both instructors and learns to gain experience from the deeper learning in the classroom. However, the learners need to take time for preparation before class and practicing in class regularly and learn on their own.

The effectiveness of the flipped classroom results from the article [11] was discovered that this learning system can include more content than the traditional lecture because of design and problem-solving scores for the engineering course. Learners' scores in the flipped classroom are higher than learners' scores in the lecture-based class only, when the types of questions are separated into problem solving or no), although learners in both groups were similar in overall scores with insignificant differences. Moreover, this paper also analyzed the student perception which expressed that most of learners like to study in the flipped classroom learning system, including using teaching online videos, while over half of the learners in the lecture class want to practice more. In the same way of the research [10], the learners must take more time to learn by themselves, though doing case studies and projects for in-class time flipped classroom learning can help the learners to understand the content deeply.

Other research [12-14] found that the flipped classroom learning has the positive impacts on the educational system and has higher performance than the traditional learning in different issues. The paper [12] adopt the flipped classroom learning to the first-year vocational college students in Taiwan. Although there is no significant effect of the flipped classroom on the certification examination of learners, the face survey interviews of learners can be concluded that using a flipped classroom to enhance student engagement can promote active learning activities both inside and outside of the classroom. In addition, the finding guided that group activities are the main goal of the in-class portion of the flipped classroom, but activities of individual learner can be helpful for persons who need more individual reflective time to learn. The study [13] compared the effectiveness of flipped classroom learning with that of traditional lecture-based learning for selected topics in the college level technology course. The results of a learner survey revealed that the learners awesomely favored use of the flipped classroom teaching style because this system supports collaboration and hands-on activities learning for solving problem during the class time. Additionally, the flipped classroom approach made a statistically significant difference to the self-efficacy in a technology integration classroom and made learners' confidence to apply knowledge more.

Another research [14] used surveying and interviewing to identify the performance of the flipped classroom learning. The results discovered that the flipped classroom can engage the learners to study more in class, and can lead to more perseverance for learning. Learners of the flipped classroom also were successful for better in productivity and the more responsibility of the learner because of various learning activities, e.g. watching videos, doing exercises and homework. Consequently, there are interesting remark points that the flipped classroom learning, focusing on learner-centered can make the learners to be satisfied with learning and teaching style. Unfortunately, the learners, who cannot be adapted to the flipped classroom learning system, will be faced with trouble and their grades may be decreased as well.

The concept and practical cases of related research [2-14] has given all benefits of the flipped classroom learning. Meanwhile, a major limitation is that the learners must need time to study on their own outside the classroom and preparing before class. Additionally, collecting suggestions of the learners in OOAD course in the previous academic year explored the learners thought that they should achieve the learning objectives of the course and can be applied to real world cases, when they have time to train the critical thinking skill for analyzing and design more besides learning object-oriented design principles, including trying analysis and design thinking on practical issue problems. These are the reasons why the objectives of this proposed research extending the research [1] is to analyze the results of using the flipped classroom learning system in OOAD course. The research [1] studied on the learners' satisfaction of this learning system and discovered the advantages and limitations of the flipped classroom learning classroom. This proposed research analyzed and reported the learners' behavior with learner opinions in detail to confirm the improvement of the learning intention and critical thinking skills of learners. Moreover, the effectiveness of the flipped classroom learning was evaluated by a comparison of learners' total scores and term project scores in two different academic years, which OOAD course in the current academic year has applied the flipped classroom learning, while in the previous academic year used the traditional lecture-based learning. Furthermore, learners' opinions about a workshop activity and the suitability of the flipped classroom learning for OOAD courses were surveyed and made a conclusion.

3. Research Methodology

3.1. Research questions

The researcher tried out the flipped classroom learning system to analyze the learners' behavior and opinions to find out whether this system encourage learning intention and critical thinking skill or not, including suitability for OOAD course. Moreover, learners' scores of the flipped classroom learning were compared to those of the traditional lecture-based learning to evaluate the effectiveness of this learning system. Furthermore, problems or obstacles in both instructor and learner views were reported for the challenges of using the flipped classroom in learning and teaching management system.

3.2. Participants and Research Instruments

There was collecting data from 20 students enrolled in object-oriented analysis and design course which were three-year and four-year undergraduate students in Chulalongkorn University. The accumulated grade point average (GPAX) of learners is displayed in Table 1. The learners' behavior, both in-class time and out-of-class time were collected for observation and questionnaire survey. Consequently, the class attendance scores, in-class assignment scores, class preparing scores, individual academic exam scores and the term project scores of the student groups were focused to analyze the effectiveness of the flipped classroom learning system. Additionally, the answers of a questionnaire about the flipped classroom learning were summarized to gain information on learners' attitudes and to conclude the results of the flipped classroom learning system.

Table 1. The Accumulated grade point average of learners

Accumulated grade point average (GPAX)	*Learner (persons)*	*Ratio of all learners*
< 2.00	0	0%
2.00 - 2.50	3	15%
2.50 - 3.00	10	50%
3.00 - 3.50	4	20%
> 3.50	3	15%

3.3. Course Activities

The course activities of the flipped classroom learning system in this proposed research are explained in Figure 1. There were data collecting of the learners' behavior, both in-class time in the classroom by observations and out-of-class time outside the classroom by a survey. The examples of course activities were watching lecture videos, doing exercises, and answering the questions about the subjects. In the classroom, there were group activities about case studies or interesting topics related to subject contents, such as discussions and presentations. Moreover, the progress of term project continued to be processed and reviewed by another group learners and the instructors in the classroom (in-class time). Furthermore, the learners gave some presentations both pair and group work for subject contents or term project works in a few weeks.

In addition, a workshop activity about analysis and design in object-oriented view was added into this course in the current academic year. The workshop issue is a real-world problem about environmental problem-solving arranged outside the classroom. All learner groups had to gather requirements, analyze and design the system to solve environmental issues.

3.4. Course Materials

The teaching materials are Thai lecture videos and English documents, e.g. PowerPoint slides created by the instructors. In addition, some online multimedia, contents, or documents which related to the subject lessons can included in the course materials, such as YouTube videos, online documents on websites, pdf file, and other online knowledge resources. Table 2 shows the number of lecture videos and their length in each week.

3.5. Course Environments

CourseVille [12], which developed by Faculty of Engineering, Chulalongkorn University, is the Learning Management System (LMS) for this OOAD course.

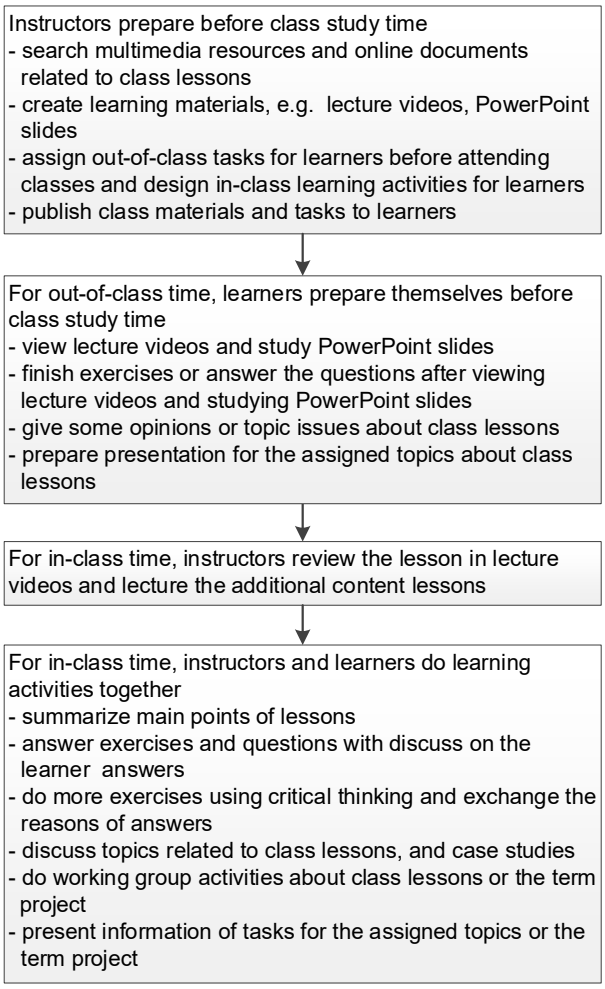


Figure 1. Course activities with the flipped classroom learning system.

Table 2 The features of lecture videos.

Week	*Number of episodes*	*Length per episode (minutes)*	*Total length (minutes)*
1	-	-	-
2	1	5.49	5.49
3	2	23.46 26.09	49.55
4	1	22.15	22.15
5	2	7.18 9.14	16.32
6	1	17.29	17.29
7	Midterm Exam		
8	-	-	-
9	2	14.16 13.09	27.25
10	1	10.15	10.15
11	1	11.46	11.46
12	-	-	-
13	1	12.05	12.05
14	-	-	-
15-16	Final Exam		

All course data on CourseVille are be stored in the cloud. Additionally, some activities can connect to Facebook to engage learners for acknowledging learning activities. For examples, CourseVille are logged by Facebook account, and links courses to the Facebook Group. General LMS services are also provided by CourseVille as follows:

- Multimedia course materials (e.g. lecture videos, video files, PowerPoint slides, pdf file documents, YouTube link) can be uploaded by instructors and can be downloaded or accessed by learners.
- News and announcements about the course can be announced to learners by instructors and can be gotten by learners.
- All teaching materials and tasks published by the instructors, can be acknowledged to the learners via the system which link to Facebook.
- Instructors and learners can raise questions and post topics to exchange their own ideas or opinions though the system.
- Instructors can assign tasks or exercises, while learners can do or answer online by the system. In addition, all learner submitted tasks can be scored online with the system by instructors
- Instructors and learners can check and record the class attendance of learners.
- Instructors can use some automatic classroom tools on the system, e.g. student randomizer, countdown timer, and Instant Q (lets instructors to create simple questions/polls to which students can respond suddenly).

4. Research Results

The proposed research experimented on the instructor and learners of OOAD course, and collected data from 20 students enrolled in this course. The results of using the flipped classroom learning in OOAD course are separated into five parts.

- The results of analysis and study on learners' behavior when using the flipped classroom learning were the analysis of learners' preparation before class, learners' class attendance, including the assignment and task submission.
- The results of effectiveness evaluation of using the flipped classroom learning were focused on academic exam scores by knowledge assessment, the term project scores which were compared to the behavior of learners, i.e. class attending, lecture video viewing and PowerPoint slide learning before class, and finishing all tasks through the LMS.
- The results of analysis and study on the learners' opinions about the flipped classroom learning and the workshop activity were summarized from the questionnaire about the flipped classroom learning with a workshop activity.
- The report of problems or obstacles in the flipped classroom learning and teaching management
- The summary of the suitability of the flipped classroom learning in OOAD course.

4.1. The Results of Analysis and Study on Learners' Behavior

In this research, interesting learners' behavior is separated into 3 main activities: preparing before class (watching lecture video and studying class documents), attending class, and submitting the assignments in class and after class). These data were collected from observations and behavior survey questions.

The learners' behavior of watching lecture videos and studying class documents before class are shown in Table 3. According to

learners' behavior survey results, while most of the learners (75% of all learners) watched almost of lecture videos (80%-100% of all videos) before class, nobody studied class documents before class more than 80% of all documents. Additionally, 45% and 35% of all learners studied the documents about 20%– 40% and 40%–60%, respectively. When the number of learners who watched lecture videos were compared to those who studied class documents, Figure 2 expressed that most of the learners from all 20 students watched the video 80-100% of all videos, but study class documents about 20-60% of all documents. The x-axis of bar graph represents the percentage of the number of watched lecture videos to the number of total lecture videos compared to the percentage of the number of class documents to the number of total class documents, and the y-axis denotes the number of students. Consequently, there were only 20% of all learners who usually study documents during video viewing. Therefore, most of the learners give precedence to watching the video more than study text documents before classes. In addition, the results concluded that most learners did not realize the importance of studying the document during the video viewing.

Table 3. Behavior of watching lecture videos and studying documents before class.

Behavior of watching lecture videos and studying documents before class		*Learner (persons)*	*Ratio of all learners*
% of the number of watched lecture videos to the number of total lecture videos	0% - 20%	-	0%
	20% - 40%	-	0%
	40% - 60%	2	10%
	60% - 80%	3	15%
	80% - 100%	15	75%
% of the number of class documents to the number of total class documents	0% - 20%	1	5%
	20% - 40%	9	45%
	40% - 60%	7	35%
	60% - 80%	3	15%
	80% - 100%	-	0%
% of the number of watched lecture videos with studying the document to the number of total lecture videos	0% - 20%	3	15%
	20% - 40%	2	10%
	40% - 60%	6	30%
	60% - 80%	5	25%
	80% - 100%	4	20%

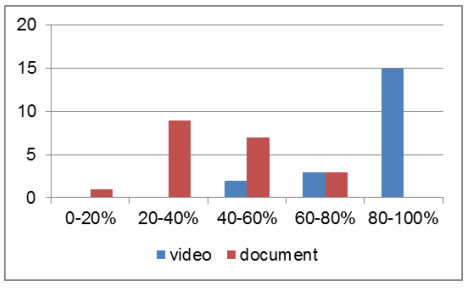


Figure 2. The number of learners watching lecture videos compare to those who studied class documents.

Moreover, the behavior analysis of learners' video watching as Table 4 found that approximately 50% of learners focused to view and listen intently to the whole part of the video. Additionally, 30% of learners view and listen through the end without serious study. Referring to the lecture videos during each week on Table 2, Table 5 also demonstrates the number of learners who did not watch lecture videos, and watched lecture videos through the end, including viewing only some parts of the lecture videos. Unfortunately, there were few learners not informing viewing types during some weeks, so this information is "No record".

Table 4. The characteristics of watching lecture videos before class.

The characteristics of watching lecture videos	*Learner (persons)*	*Ratio of all learners*
only listen through the end	1	5%
only listen intently some part of the video	-	0%
view and listen through the end	6	30%
view and listen intently some part of the video	2	10%
view and listen intently the whole part of the video	11	55%

Table 5. The video viewing types of learners for each week.

Week	*Not view*	*View the end*	*View not end*	*No record*	*Total (persons)*
1	-	-	-	-	-
2	3	6	7	4	20
3	4	5	8	3	20
4	-	12	6	2	20
5	1	18	-	1	20
6	1	15	4	-	20
7	Midterm Exam				
8	-	-	-	-	-
9	3	8	6	3	20
10	-	17	1	2	20
11	2	14	1	3	20
12	-	-	-	-	-
13	-	18	-	2	20
14	-	-	-	-	-
15-16	Final Exam				

According to the features of the lecture videos on Table 1 and learner's viewing behavior on Table 5, during the week that had only a single video episode, learners would watch through the end more than the week that had more video episodes. In addition, learners would obviously watch less through the end during the week with the length of all videos is about or more than 20 minutes. Then, if there is more than one video episode during the week, more learners would watch through the end of the shorter length of the episode. In the same way, if the length of a video episode is about or less than 10 minutes, learners would watch through the end of every episode more.

Furthermore, the results from the questionnaire surveyed learners' time duration preparing for class found that the preparation time by learners 55% of all learners around about 12-24 hours before class as shown in Table 6, and nobody also do prepare more than 1 day. Another result discovered that the popular time to prepare for the class was during the late evening to midnight as displayed in Table 7. It was not surprising that most learners preparing themselves in the night before the morning classes.

Table 6. The preparation time before class.

Time duration	*Learner (persons)*	*Ratio of all learners*
before class less than 3 hours	4	20%
before class 3 - 12 hours	4	20%
before class 12 - 24 hours	11	55%
before class 1 day	1	5%
before class more than 1-2 days	-	0%

Table 7. The time of preparing for class.

Time	*Learner (persons)*	*Ratio of all learners*
0.00 - 4.00 a.m.	2	10%
4.00 - 8.00 a.m.	2	10%
8.00 - 12.00 a.m.	-	0%
0.00 - 4.00 p.m.	-	0%
4.00 - 8.00 p.m.	5	25%
8.00 - 12.00 p.m.	11	55%

Table 8. The class attendance of individual learners.

Learners no.	*Not attend*	*In time*	*Late*	*No record*	*Total (weeks)*
1	-	12	-	-	12
2	2	9	1	-	12
3	-	12	-	-	12
4	4	7	1	-	12
5	-	11	-	1	12
6	1	5	6	-	12
7	-	5	5	2	12
8	1	-	11	-	12
9	1	11	-	-	12
10	1	6	5	-	12
11	-	9	3	-	12
12	-	12	-	-	12
13	-	12	-	-	12
14	4	6	1	1	12
15	2	8	2	-	12
16	-	12	-	-	12
17	-	11	1	-	12
18	7	2	2	1	12
19	-	6	4	2	12
20	4	3	5	-	12
Total	11.25%	65%	19.58%	2.92%	100%

Table 9. The time of coming in class.

Time	*Learner (persons)*	*Ratio of all learners*
8.00 - 8.30 a.m.	4	20%
8.30 - 9.00 a.m.	8	40%
9.00 - 9.30 a.m.	6	30%
9.30 - 10.00 a.m.	1	5%
10.00 - 10.30 a.m.	-	0%
10.30 - 11.00 a.m.	1	5%

To observe learners' behavior in class, attendances in the classroom of individual learners were checked for 12 weeks displayed in Table 8. The OOAD course class time was 3 hours per week and started at 8.00 a.m. on Monday. The learners who came in the classroom before 9.00 a.m. were on time, otherwise they were late. The results found that 65% of all attendance checking are in time because most learners attended classes before 9.00 a.m., whether they missed the class. However, there was one learner always late for class, and another learner usually missed the class.

In addition, the learners' common time coming in class is illustrated in Table 9. Based on the time attendance, there are 60% of all learners attending classes before 9 a.m. and participating the class activities in time. Although there were a group of late participants (30% of all learners), they did not late more than a half an hour. Therefore, most learners realize that it is important to attend class and practice themselves in the flipped classroom learning.

By the way, all assignment submissions of learners recorded on CourseVille were composed of exercises or topic discussions in class and individual assignments after class. Table 10 presented the number of learners who submitted the assignments, and the rest of them who did not submit the assignments. The results show that almost all the learners submitted assigned works every piece, except few learners.

Table 10. Individual class attendance.

Assignment no.	*Submission*	*Not submission*	*Total (persons)*
1	17	3	20
2	13	7	20
3	13	7	20
4	19	1	20
5	20	-	20
6	20	-	20
7	17	3	20
8	16	4	20
9	20	-	20
10	16	4	20
Total	85.5%	14.5%	100%

In conclusion, all learners' behavior expressed that most learners gave corporation for managing the flipped classroom learning. They took their time to watch lecture videos for preparing before class and attended class regularly. Although few learners missed the class or did not study before class, they did their best when joined the group activities in class. In addition, learners' preparation for class by themselves, attending class to do in-class time activities, and doing all individual assignments, were critical thinking skill training. Therefore, the flipped classroom learning can stimulate learners' intention and encourage critical thinking skill of learners.

4.2. The Results of Effectiveness Evaluation of Using the Flipped Classroom Learning

The results of effectiveness evaluation of using the flipped classroom learning are separated into 2 main parts. The first part is the comparison among evaluation scores of course exams and activities that are the midterm exam, the final exam, the term project, the assignments, the class attendance, and the video viewing before class. The second part is the comparison between evaluation scores of learners in the current academic year with the flipped classroom learning, and in the previous academic year with traditional lecture-based learning.

There were comparisons between all scores of individual learners which consisted of the assessment scores of academic knowledge of each learner (the total scores of the midterm and the final academic exams), and the scores of each course activity in the flipped classroom learning system that were watching lecture video in preparation before in-class time, the class attendance for joining group activities, and individual assignment or task submission for in-class time). All scores of individual learners were inconsistent and difficult to find their relations shown in Figure 3. The x-axis denotes the individual student number (1 - 20) and the y-axis represents the scores of learners composed of the midterm and final exam scores (out of 45 points), the class attendance scores (out of 12 points), the individual assignment scores (out of 10 points), and the lecture video viewing (out of 9 points).

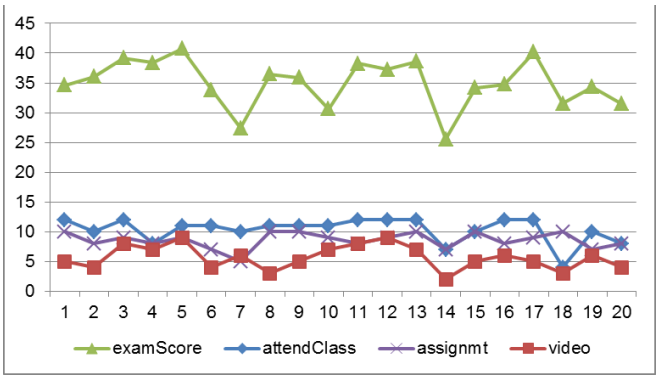


Figure 3. The exam scores of individual learners compare to each course activity.

Meanwhile, it was found that almost learners' exam scores were consistent with the summation scores of all course activities in the flipped classroom learning, when all course activity scores, which were watching lecture videos, attending class, and submitting assignments, were combined into the total scores of course activities.

The comparison result has been revealed in Figure 4, which the x-axis and y-axis mean the same parameters of Figure 3. The two-line graphs in Figure 4 demonstrated that learners, who did not prepare before class by watching lecture videos, did not join in-class learning activities by attending the class, and did not submit their works, would have low academic exam scores. And there were only a few learners who did all course activities, but still got low academic exam scores.

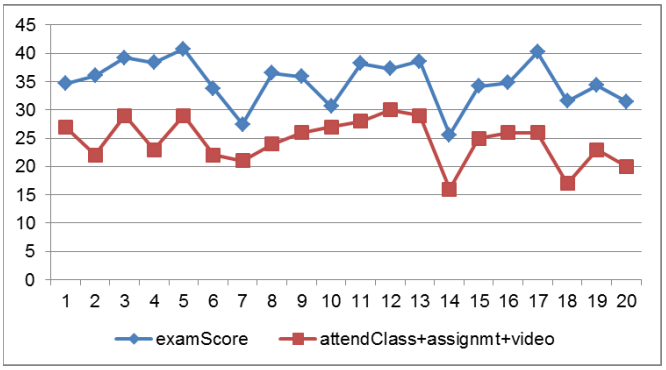


Figure 4. The academic exam scores of individual learners compared to the total scores of three course activites.

Moreover, the correlation analysis of academic exam scores with total scores of three course activities was performed to confirm the relationship between learners' academic exam scores and the total scores of learners' course activities. The analysis result displays that the academic exam scores are significantly related to the summation of all three course activities as in Figure 5. Therefore, the result has been concluded that the learner academic knowledge is not based on only one ability or one skill trained by only one activity but generated from the integrity of all learning and training activities. Knowledge improvement of individual learners requires studying and practicing all activities that are preparation before studying in-class time, paying attention to the discussion or group activities of in-class time, and training on course exercises or assignments.

Correlations

		examScore	attendClass +assignmt+ video
examScore	Pearson Correlation	1	.731**
	Sig. (2-tailed)		.000
	N	20	20
attendClass+assignmt +video	Pearson Correlation	.731**	1
	Sig. (2-tailed)	.000	
	N	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 5. The correlation between the exam scores of individual learners and the total scores of three course activities.

Furthermore, the flipped classroom learning increases time in class for discussing related topics, practicing on analysis and design skills, and doing the term project in the classroom under the guidance and management of the instructors. Therefore, the comparison results as in Figure 6 expresses that the term project scores of individual learners (out of 30 points) were related to their class attendance scores, except a few learners having the good term project score without attending class regularly. One cause of these scores is the term project was a group work, so the other members of the group could operate and review the work of absent members. The x-axis and y-axis of Figure 6 were the same meaning as Figure 3 and Figure 4. The correlation analysis of the term project and the class attendance was also displayed in Figure 7.

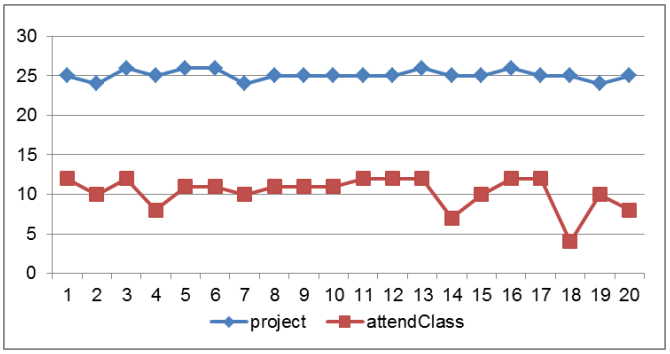


Figure 6. The term project score of individual learners compare to the class attendance

Correlations

			project	attendClass
Spearman's rho	project	Correlation Coefficient	1.000	.469*
		Sig. (2-tailed)		.037
		N	20	20
	attendClass	Correlation Coefficient	.469*	1.000
		Sig. (2-tailed)	.037	
		N	20	20

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 7. The correlation between the term project scores of individual learners and the class attendance

To evaluate the performance of the flipped classroom learning, the total scores of grading and the term project scores of learners in the current academic year were compared to those of learners in the previous academic year by the independent t-test. These scores

are also used to measure critical thinking skills of learners in this research. Figure 8 shows the term project scores and the total scores of the current academic year with the flipped classroom learning (the table at the top of the figure) and the previous academic year with the traditional learning (the table at the bottom of the figure).

Descriptive Statistics

Current	N	Minimum	Maximum	Mean	Std. Deviation
project	20	24.00	26.00	25.1000	.64072
TotalScores	20	68.83	95.10	85.2350	7.29223

Descriptive Statistics

Previous	N	Minimum	Maximum	Mean	Std. Deviation
project	10	22.00	26.00	24.2000	1.39841
TotalScores	10	73.07	88.07	80.1610	5.80154

Figure 8. the term project scores and the total scores of both academic years.

The average total scores of both academic years and the independent t-test results are demonstrated in Figure 9. The result found that the average total scores of learners in the current academic year were higher than those of learners in the previous academic year.

Group Statistics

year		N	Mean	Std. Deviation	Std. Error Mean
Total Scores	current	20	85.2350	7.29223	1.63059
	previous	10	80.1610	5.80154	1.83461

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Total Scores	Equal variances assumed	.264	.612	1.913	28	.066	5.07400	2.65244

Figure 9. The t-test results of the total scores of students in current academic year and the previous academic year

According to Figure 9, the p-value of Levene's test was greater than 0.05 (Sig. 0.612 > α), that meant the variances of two learner groups (students in the current and the previous academic years) are equal. Additionally, the p-value of t-test was less than 0.05 (Sig. 0.066/2 < α), then there was a mean difference between total scores of learners in the current academic year and the previous academic year. Therefore, the average total scores of learners in the flipped classroom learning were significantly higher than those in the traditional learning.

Another result was that the average term project scores of learners in the current academic year were higher than those of learners in the previous academic year. The average term project scores of both academic years and the independent t-test results are explained in Figure 10. Like the total scores, the variances of term project scores were equal across the two learner groups because the p-value of Levene's test was greater than 0.05 (Sig. 0.001 > α). Then, the t-test result displayed that the p-value of t-test is less than 0.05 (Sig. 0.021/2 < α), thus there was a mean difference between term project scores of students in the current academic year and the previous academic year. That meant the average term project scores of learners in the flipped classroom learning were a little bit significantly higher than those in the traditional lecture-based learning.

Group Statistics

year		N	Mean	Std. Deviation	Std. Error Mean
project	current	20	25.1000	.64072	.14327
	previous	10	24.2000	1.39841	.44222

Independent Samples Test

		for Equality of		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
project	Equal variances assumed	15.108	.001	2.440	28	.021	.90000	.36888

Figure 10. The t-test results of the term project scores of students in current academic year and the previous academic year

Although there was not much difference in the average term project scores, the effectiveness of the flipped classroom learning has been higher than the traditional learning. The reason is that the number of students in the current academic year were more than those in the previous academic years and all learner groups did the term project with the good quality. Therefore, the result can be concluded that the flipped classroom learning helped to encourage the critical thinking skill for the learners.

4.3. The Results of Analysis and Study on the Learners' Opinions

The learners' opinions were collected by the questionnaire about the features of the lecture video and all course activities in the flipped classroom learning, including a workshop activity. All interesting course activities are watching lecture videos and studying lesson documents before class (out-of-class time), working in groups and discussing together on case studies and related topics about lessons, answering instructors' questions following to discuss on learners' answers (in-class time), and doing term project with the discussion (in-class time). In addition to the flipped classroom activities, the learner attitudes about this learning system suitability for the OOAD course were summarized from the survey answers.

The opinions about the features of the teaching videos were that 60% of learners (12 students) commented the lecture video is able to have more than one episode per week, while the rest of them (40 % of learners or 8 students) supposed that it should have one episode per week only. When analyzing the length of the video, 85% of learners (17 students) thought that it should be no longer than 20 minutes or less than 10 minutes as detailed in Table 11. In addition, a comparison between the lecture video features and the number of learners who watch through the end is demonstrated in Table 12.

According to Table 12, the lecture video features were analyzed by using the number of learners who watching the video through the end. The opinion of the learners discovered that most of the learners would watch through intentionally, if the length of

one lecture video is about 10 minutes. Meanwhile, learners would noticeably watch less through the end, if the length of lecture videos is more than 20 minutes. The analysis results of the number of episodes and the length of the video per one episode are also consistent with the number of learners who watching the video until the end of each week.

Table 11. The learner opinion about the appropriate length for one episode of lecture video.

The length for one episode of video	*Learner (persons)*	*Ratio of all learners*
less than 10 min.	8	40%
10 - 20 min.	9	45%
20 - 30 min.	2	10%
more than 30 min.	1	5%

Table 12. Comparison between the lecture video features and the number of learners who watch through the end.

Week	*Number of episodes*	*Total length of lecture videos (minutes)*	*Number of learners viewing through the end (persons)*
1	-	-	-
2	1	5.49	6
3	2	49.55	5
4	1	22.15	12
5	2	16.32	18
6	1	17.29	15
7	Midterm Exam		
8	-	-	-
9	2	27.25	8
10	1	10.15	17
11	1	11.46	14
12	-	-	-
13	1	12.05	18
14	-	-	-
15-16	Final Exam		

To ensure the performance of the flipped classroom learning, there were seven survey questions about the impacts of the learning course activities in the flipped classroom learning for assessing learners' opinions.

The effectiveness evaluation results are the average attitude scores from all learners' opinions that is classified in five levels:

1 = A disagree 2 = B quite disagree 3 = C partly agree
4 = D quite agree 5 = E agree

Figure 11 shows the number of learners who commented on each survey question and the conclusions of the learners' opinions that agree or disagree on each topic are displayed as follows:

- (1) Watching lecture videos before in-class time helps learners to understand the subjects more easily. Mean score = 4.25
- (2) Studying lesson documents before in-class time helps learners to understand the subjects more easily. Mean score = 3.90
- (3) Case study and group discussion in class make learners to understand the subjects deeply. Mean score = 4.00
- (4) In-class working group and discussing together make learners to understand the subjects deeply. Mean score = 3.85

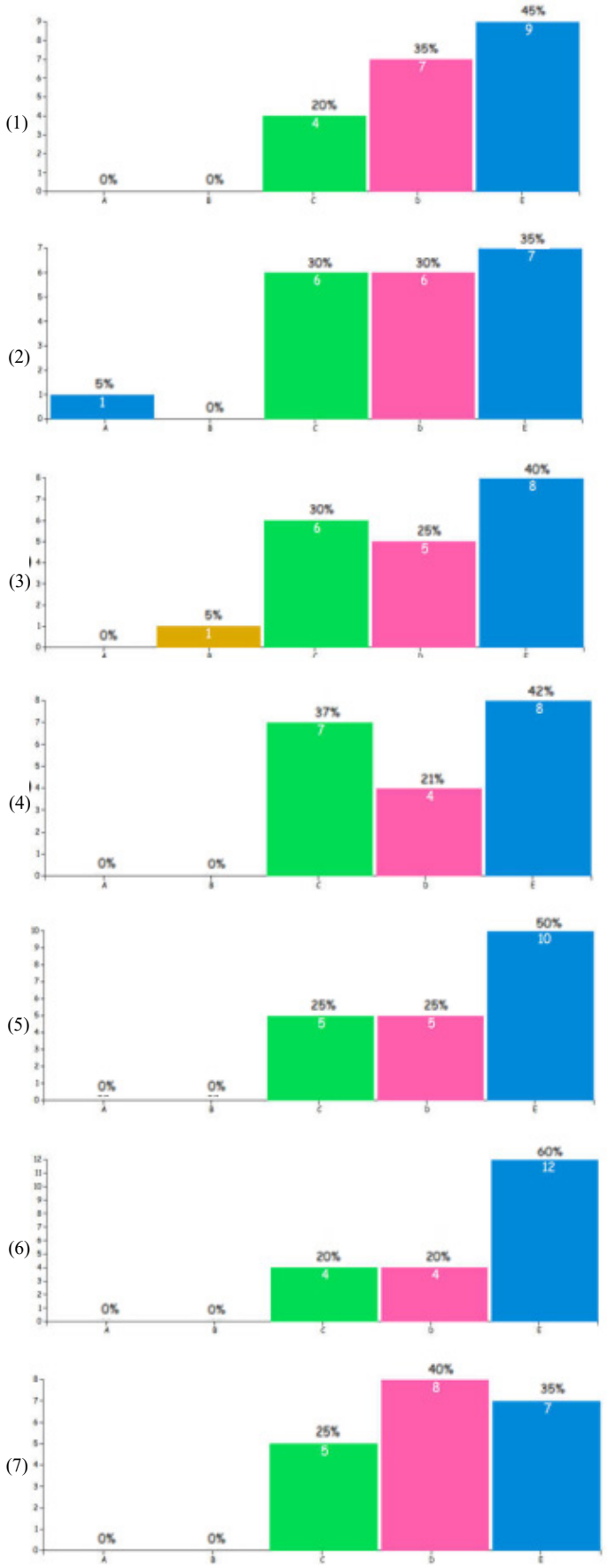


Figure 11. The learner opinions about the flipped classroom learning activities

- (5) Doing term project in class study time with discussion together helps learners to develop more quality of the term project. Mean score = 4.25
- (6) Answering questions with discussion on the learners' answers help the learner to improve their knowledge. Mean score = 4.40
- (7) The flipped classroom learning helps to develop critical thinking, analysis and design skills more. Mean score = 4.10

According to the bar graphs in Figure 11, most of the learners agree that study preparation before in-class time (out-of-class time) and in-class time activities in the flipped classroom learning make learners to understand the lesson contents easier and deeper, develop a better quality of term projects, and improve their own knowledge. In conclusion, most of the learners are agreed and quit agree that the flipped classroom learning helps to develop critical thinking skill, including analysis and designing skills better than the traditional lecture-based learning.

By the way, the opinion assessment of the learners about the object-oriented analysis and design workshop was evaluated by the questionnaire on five levels like effectiveness assessment. The average scores of learners' opinion for each topic are shown as follows:

- (1) You have learned how to analyze with critical thinking the overview of preserving the underwater environment. Mean score = 3.84
- (2) You have learned to identify the problem of coral damage. Mean score = 4.11
- (3) You understand the causes of problem in coral damage. Mean score = 4.26
- (4) You have gotten experiences in interviewing to gather requirements for problem solving about coral damage. Mean score = 3.84
- (5) You have gotten experiences in the analysis and design for the real-world problem. Mean score = 4.16
- (6) You have practiced on real world problem solving. Mean score = 4.05
- (7) You understand the subject contents more by joining this workshop activity. Mean score =3.95
- (8) You have trained on group working to create good quality works. Mean score =4.16

Referring to the average scores of each survey question of the workshop activity, most of the learners agreed that learners have learned how to analyze with critical thinking and identify the problem, including to find out and understand the causes of the problem. Moreover, the learners had gotten experiences in gathering requirements, analysis and design, and solving the real-world case study. Furthermore, most learners can understand the subjects deeply by doing workshop activities and practice themselves in working groups for creating good quality works.

4.4. The Report of Problems or Obstacles

There were two different question types of data collection for problems and obstacles of the flipped classroom learning. The first question with answer choices were asked for the most important problem or obstacle in the individual learners' attitude. And the second question with the short answers was independent writing for the question "What are your difficulties or obstacles in the flipped classroom learning?".

According to the selected answers by learners, 55% of the learners (11 students) thought that the most important problem or obstacle in the flipped classroom learning system is habit and self-responsibility of individual learners. Meanwhile, some learners (4 students) are not able to understand the content in the lecture video or the PowerPoint slide that are the most important problem for them in this learning system. However, there are two students told that their most important problem is no time to study the preparation before class, and only one student who thought that unfriendly LMS, or no time to attend class, or conflicts on discussions or group works is the most important problem.

For the independent writing answers, there were five students left empty and three students wrote "no problem". The comments of the rest students were grouped as follows:

- Sometimes there is no time for study preparation before in-class time, so it caused the less understanding and not learn much.
- The instructors upload some lecture videos before class study time less than one week, if we are busy, we cannot watch them in advance.
- Cannot understand all the content, and need to study more or ask more in class study time.
- Cannot get the answers to any questions.
- Need time to adjust to this learning system because of the unfamiliar learning system.
- Viewing the lecture video is very troublesome because of the too small screen.
- Get bored with watching the lecture videos through the end because some lecture videos are too long.
- Need the instructors to upload and publish all lecture videos at the beginning course period because we can totally learn when we have a spare time.
- Problems with the internet connection lead to other problems. For example, sometimes cannot display the lecture videos, cannot access to the LMS, and miss the class announcements

Other problems or obstacles of the flipped classroom learning in the instructor view are displayed as follows:

- The learners lack time to watch the video or study the lesson document before in-class time, and if there are no scores for self-study preparing activities, the learners may not concentrate on viewing the lecture video or learning class documents in advance.
- There is insufficient time for the instructors to prepare learning materials, such as lecture videos and PowerPoint slides for the whole subject contents of the course in advance, so sometimes they has been published before classes less than one week.

- There is the variant of the lesson contents so some lessons in some weeks cannot be understood by viewing the lecture videos with a length of shorter than 20 minutes.

4.5. The Summary of Suitability of the Flipped Classroom Learning

The assessment results of the survey answers from the questionnaire were analyzed and the mean score of the appropriateness of the flipped classroom learning system was calculated to summarize the suitability of the flipped classroom learning for this course. The rating given by the learners on five levels as below:

1 = A inappropriate 2 = B quite inappropriate
3 = C moderately appropriate 4 = D quite appropriate
5 = E appropriate

The mean score for the appropriateness of the flipped classroom learning was 4.40. The result means almost of the learners believed that the flipped classroom learning system has appropriated for this course, as detailed in Figure 12.

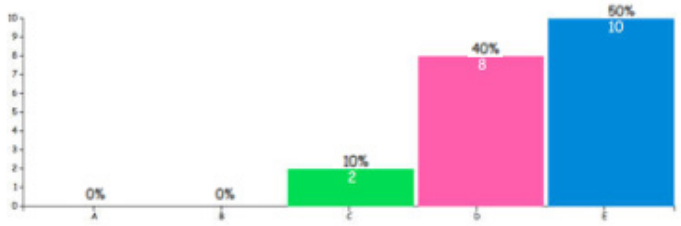


Figure 12. The number of learners who answered the suitability of the flipped classroom learning.

According to Figure 12, there were 50% of learners (10 students) and 40% of learners (8 students) thought that the flipped classroom learning has been appropriate and has been quite appropriate to the OOAD course. Therefore, these results show that the flipped classroom learning is suitable for the OOAD course with the aggregation of 90 percent of learners.

Moreover, the reasons which support their attitudes are explained as listed below:

- There can be the diversity in the thinking process in this course to define the functionality of the system and design the system. Therefore, explaining, dialoguing, questioning and answering among learners, including between learners and instructors can help in learning as well. The discussion between learners and instructors helps them to have coincident ideas.
- The content of this course is quite many subjects. The flipped classroom learning helps learners to understand the lesson before class, have time to exchange ideas and work in groups more in class.
- This course focuses on the practical so learners have to think and practice by themselves before they consult or discuss with others. In addition, learners can learn, think, ask the question and find the answer by themselves to use their own ability for learning.
- This course is not only focused on academic content, but also emphasize on practical application. It should have activities in the classroom instead of teaching the content which learners can read from documents or find information from outside resources. It also needs to discuss and practice to understand the contents because the subjects cannot be understood by only reading.
- The flipped classroom learning system is very beneficial. The learners are enabled to practice critical thinking and self-study skills. Watching the video before class makes an easier preparing and more intention. Discussion motivates the learners to be active, not sleepy in the classroom.
- If learners cannot attend classes or some contents may be difficult, they are able to study some subjects of the video viewing outside classes which may make more understanding. The flipped classroom learning system helps to save the time of teaching in lecture class duration, so the interesting subjects can be learned more.

Furthermore, the advantages of the flipped classroom learning that the learners gave opinions on their own side are listed below:

- Learners can study the lessons before class that make easier understanding when they come to learn in class. If they don't understand the lessons, or have any questions, they can ask the instructor in the classroom. Beside previous advantages, they pay more attention in class because they make understanding the contents before class.
- Learners have selves' critical thinking before class, so they can remember and understand the contents more, don't have to read much for exams.
- Learners have practiced critical thinking and done practical works under the care and guidance of the instructor.
- Learners can understand the lessons more and work better on the project because the traditional teaching system spent the whole period for lecturing, no chance to ask questions about the project, and no advisor to check whether their works are right or not.
- Learners can have the lecture video to watch after class and prepare for tests that make reviewing the lessons easily. Additionally, in the case of missing the class, learners still study from watching the video and gain some knowledge.
- Learners can discuss and exchange ideas in the class without wasting the teaching time.
- A non-stressful environment for learning in class.

In conclusion, although there are problems or obstacles in the flipped classroom learning because of limitation out-of-class time for learners preparing a lesson before class and for instructors providing teaching materials, including specifying in-class time activities, there are more benefits to use this new learning system in the OOAD course with many supported reasons by learners' opinions. Therefore, the flipped classroom learning has been suitable for the OOAD course in this research results.

5. Conclusions

Using of the flipped classroom learning in the OOAD course was studied to analyze the results of using this learning system. This research focused on the behavior of learners, learners' scores, and learner opinions to find out the appropriateness of the flipped classroom learning in this course. The results of learners' behavior

analysis were found that most learners are interested in the flipped classroom learning activities and enjoy cooperating with the instructors. In addition, the performance evaluation of the flipped classroom learning for the OOAD course can be concluded that this learning system has been better than the traditional learning system because the average total scores of grading and the average term project scores of learners in the current academic year with the flipped classroom learning were significantly higher than those in the previous academic year with the traditional learning. Therefore, using the flipped classroom learning can increase the learning outcome of the learners and can be useful for learners in the learning intention and critical thinking skill development.

The summary of learners' opinions on the flipped classroom learning with a workshop activity was uncovered that the majority of the learners concurred that the flipped classroom learning activities composed by viewing the lecture video and studying the text documents before in-class time, helps them to understand the class contents effortlessly. Moreover, group discussions of case studies, and doing a term project under the guidance of the instructors in class study time can help understand class contents well and create the better term project. Furthermore, the solutions to the assignments or the answers of exercises which are explained supporting reasons in the classroom make the learners develop their own knowledge as well.

In conclusion, most of the learners agreed that the flipped classroom learning system can help improve critical thinking skill and object-oriented analysis and design skill. The research results also pointed out that the flipped classroom learning system helps the learners to pay more attention because of preparation before class. Meanwhile, the critical thinking skills were cultivated by discussions, group activities in-class time, which can ask questions and get advice from the instructor immediately when in trouble, so it encourages learners to practice and apply the knowledge better. In addition, there were 90% of learners thought the flipped classroom learning is suitable for this course. Therefore, it can be concluded that the flipped classroom learning is suitable for the OOAD course.

6. Suggestions

In spite of the fact that the flipped classroom learning is appropriate to the OOAD course, this learning system requires readiness in both of the instructors and the learners. They should prepare for adjusting themselves to the active learning style with the useful learning management system. Instructors should be able to generate suitable learning materials with valuable knowledge, and to transfer that knowledge to learners. Moreover, learners must have abilities in self-study learning for learning and understanding course content. In addition, the learner need to concentrate on study preparation before in-class time and doing learning group in-class activities to get their knowledge and practice their skills. Furthermore, both instructors and learners should be able to create a friendly atmosphere in the classroom.

In fact, according to problems and obstacles of the flipped classroom learning in views of both instructors and learners, this learning system cannot completely be flipped the classroom style. If instructors have more time to create learning materials with the coordination between different courses in the same field, such as software development, this may increase the achievement of the flipped classroom learning because learners can arrange appropriate times for self-study preparation and cooperating group activities, while the instructor can share available learning resources in order to save time for searching and generating a few duplicate subjects in different courses.

Beside available time of learners and instructors, materials in the course which are prepared or provided by instructors cannot guarantee that all learners are able to study and understand by themselves. The reasons are that the content of each lesson has the different difficulty, and there still are differences among self-learning skill, intelligence, ability and perseverance of each learner. Additionally, the expertise of individual instructors to transfer knowledge in each lesson is different. If there are quality and standard materials, may cause the flipped classroom learning system achievement even better.

Acknowledgment

This classroom action research got funded by Learning Innovation Center, Chulalongkorn University.

References

[1] P. Pugsee, "Effects of using flipped classroom learning in object-oriented analysis and design course" in 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media), Pattaya, Thailand, 2017. https://doi.org/10.1109/UMEDIA.2017.8074130

[2] C. Brinton, M. Chiang, "Social Learning Networks: A brief survey" in 48th Annual Conference on Information Science and Systems (CISS), Princeton, NJ, USA, 2014. https://doi.org/10.1109/CISS.2014. 6814139

[3] A. Wallace, "Social learning platforms and the flipped classroom" Int. J. Inf. Educ. Technol., 4(4), 293-296, 2014. https://doi.org/10.7763/ IJIET.2014.V4.416

[4] T. Bristol, "Flipping the classroom" Teach. Learn. Nurs., 9(1), 43-46, 2014. https://doi.org/10.1016/j.teln.2013.11.002

[5] Q. Dang, D.D. Gajski, "Bringing in-class online - A hybrid solution" in 4th Interdisciplinary Engineering Design Education Conference (IEDEC), Santa Clara, CA, USA, 2014. https://doi.org/10.1109/ IEDEC.2014.6784674

[6] A. Amresh, A.R. Carberry, J. Femiani, "Evaluating the effectiveness of flipped classrooms for teaching CS1" in IEEE Frontiers in Education Conference (FIE), Oklahoma City, OK, USA, 2013. https://doi.org/ 10.1109/FIE.2013.6684923

[7] J.M.M. Ferreira, "Flipped classrooms: From concept to reality using Google Apps" in 11th International Conference on Remote Engineering and Virtual Instrumentation (REV), Porto, Portugal, 2014. https://doi.org/ 10.1109/REV.2014.6784256

[8] M. L. Carrió-Pastor, H. Skorczynska, "Collaborative learning and communication technologies in teaching business english" Procedia Soc. Behav. Sci., 178, 32-37, 2015. https://doi.org/10.1016/j.sbspro. 2015.03.142

[9] K. T. Wissman, K. A. Rawson, "Why does collaborative retrieval improve memory? Enhanced relational and item-specific processing" J. Mem. Lang, 84, 75-87, 2015. https://doi.org/10.1016/j.jml. 2015.05.003

[10] C. P. Rosiene, J. A. Rosiene, "Flipping a programming course: The good, the bad, and the ugly" in IEEE Frontiers in Education Conference (FIE), El Paso, TX, USA, 2015. https://doi.org/10.1109/FIE.2015. 7344151

[11] G.S. Mason, T.R. Shuman, K.E. Cook, "Comparing the effectiveness of an inverted classroom to a traditional classroom in an upper-division engineering course" IEEE Trans. Educ. 56(4), 430-435, 2013. https://doi.org/10.1109/TE.2013.2249066

[12] Y. Chen, L. Chen, "Effects of the flipped classroom model on student performance for vocational college students" in International Conference on Educational Innovation through Technology (EITT), Tainan, Taiwan, 2016. https://doi.org/10.1109/EITT.2016.30

[13] M. Khan, M. Ibrahim, "Flipped classroom in technology courses - impact on personal efficacy and perception based on learning style preferences" in IEEE Integrated STEM Education Conference (ISEC), Princeton, NJ, USA, 2017. https://doi.org/10.1109/ISECon.2017. 7910229

[14] A. Amiri, H. Ahrari; Z. A. Saffar; V. Akre, "The effects of classroom flip on the student learning experience: An investigative study in UAE classrooms" in International Conference on Current Trends in Information Technology (CTIT), Dubai, United Arab Emirates, 2013. https://doi.org/10.1109/CTIT.2013.6749480

[15] CourseVille. "The Social Learning Management System of Chulalongkorn University". https://www.mycourseville.com/

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 53-58 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Dc-link voltage drift compensation in a four level Double-Star Converter using redundant states via phase-shifted PWM strategy

Carlos A. Reusser*

Universidad Tecnica Federico Santa Maria, Department of Electronics, 2390123, Valparaiso, Chile

ARTICLEINFO

Article history:
Received: 07 May, 2018
Accepted: 30 June, 2018
Online: 14 July, 2018

Keywords:
Double-Star Converter
Redundant Switching States
DC-Voltage Drift

ABSTRACT

The following work introduces the use of the available Redundant switching states introduced by the double star four level converter topology, for its application in a medium voltage variable speed drive under a wide range of operation loads. Redundant switching states are obtained using multi-carrier phase-shifted pulse width modulation (PWM) strategy. The converter performance is evaluated in terms of the dc-voltage drift and the torque ripple on the load side, when compared with a classical level-shifted PWM modulation.

1 Introduction

Modern industrial high demanding processes are commonly based on medium voltage induction machines (MV-IM) AC-drives with Field Oriented Control (FOC) scheme [1]. In this field of applications, voltage source multi-level converters are the dominant topologies used in industry, rather than two-level voltage source converters (2L-VSC), in order to reduce the blocking voltage rating of power switches (semi-conductors) as well as conduction losses; harmonic distortion is also minimized ensuring a low rate of torque ripple [2].

Commonly adopted converter topologies for industrial applications are based on cascade H-Bridge (CHB), neutral point clamped (NPC) and flying capacitor (FC) converters [3, 4, 5, 6]. This configurations can be considered as the today's industry standard, because they have reached heir technological maturity proven during the past couple of decades in high demanding industrial processes. However, due the fact that variable speed drives operate within a wide range load profile, most of the time at a fraction of their nominal load, multi-level converters have to deal with this fact, operating in their low modulation index region.

Nowadays, level-shifted PWM (LS-PWM) strategy has become a standard as modulation strategy for multi-level converters. However this modulation technique has a poor performance at low modulation indexes, resulting in dc-link voltage unbalance and the consequent torque ripple in the mechanical drive [7].

The recent diversification of the power electronics market, has open a new scenario for medium voltage converter topologies that can compete with today's industrial standards, improving the disadvantages of classical converter topologies and keeping their advantages (medium voltage capability, low Total Harmonic Distortion (THD) and low semi-conductor stress) [8].

Following this motivation, in this work a double star four level medium voltage converter topology [9] is introduced for its application in medium-voltage AC-drives. However the topology offers many advantages over commonly used industrial standards, it shows the same problem introduced by level-shifted PWM at low modulation index, as reported in [9].

The main contribution of the of the present work, is referred to the use of the redundant switching states available in the double star topology, to deal with the dc-link voltage unbalance, in low modulation index operation. Redundant switching states are synthesized using a phase-shifted PWM (PS-PWM) scheme. The performance of the proposed solution, is analysed via simulation results using an induction machine Filed Oriented Control drive as industrial load.

2 Double Star Multi-level Converter Topology

The double-star converter (DSC) topology is based on a three-level NPC (3L-NPC) configuration, but with the absence of clamping diodes or switches (as in the case of the ANPC) [9]. The converter front-end is arranged

*Corresponding Author: Carlos A. Reusser; Av. Espana 1680, Valparaiso, Chile; carlos.reusser@usm.cl

https://dx.doi.org/10.25046/aj030407

with an 18-pulse diode rectifier, fed from the main grid via a multi-winding zig-zag phase-shifting transformer so that a shifting in -20, 0 and 20 degrees, is achieved. Each diode rectifier is connected to a dc-link capacitor (each level) and are also interconnected in series to each other as presented in Fig. 1.

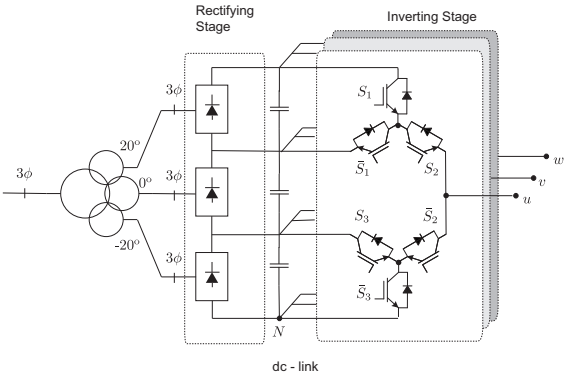


Figure 1: Double Star converter topology

The most remarkable features of this topology are:

1. 6 active switches per phase like the three-level Active Neutral Point Clamped converter (3L-ANPC).

2. 4 voltage levels; one more than the 3L-ANPC and using less number of active switches.

3. no need of clamping diodes or switches (ANPC and NPC topologies), isolated DC-sources (Cascaded H-bridge CHB) or flying capacitors.

The absence of clamping devices makes that each capacitor is in a floating condition, respect to ground, such as in a three level Flying Capacitor converter (3L-FC). The converter fundamental cell (per leg) is presented in Fig. 2 and its corresponding allowed switching states are listed in Table 1

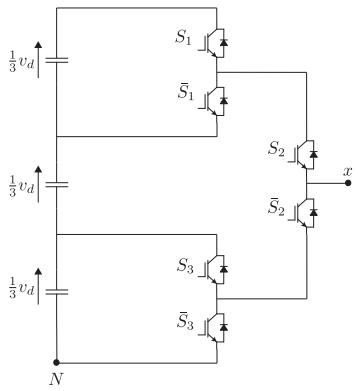


Figure 2: Fundamental converter cell

Table 1: Switch stages per leg (fundamental cell)

#	S_1	S_2	S_3	v_{xN}
1	0	0	0	0
2	0	0	1	$\frac{1}{3}v_d$
3	0	1	0	$\frac{2}{3}v_d$
4	0	1	1	$\frac{2}{3}v_d$
5	1	0	0	0
6	1	0	1	$\frac{1}{3}v_d$
7	1	1	0	v_d
8	1	1	1	v_d

Pulse width modulation techniques for multi-level converters, such as level-shifted PWM (LS-PWM) and phase-shifted PWM (PS-PWM), are based on the use of multiple carriers, as an extension of two-level PWM methods. Level-shifted PWM (LS-PWM) strategy has become a very popular multi-level modulation technique, because it is suitable for any multi-level converter topology and presents low harmonic distortion [10, 11]. The corresponding output voltage is synthesized as given in eqs. (1) - (5)

$$v_{xn} = \frac{1}{3}\, v_d\, (S_1 + S_2 + S_3) \quad \forall\, x \in [a, b, c] \tag{1}$$

$$S_1 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c1}| \\ 0 & |u^*_{xn}| < |u_{c1}| \end{cases} \tag{2}$$

$$S_2 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c2}| \\ 0 & |u^*_{xn}| < |u_{c2}| \end{cases} \tag{3}$$

$$S_3 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c3}| \\ 0 & |u^*_{xn}| < |u_{c3}| \end{cases} \tag{4}$$

where u^*_{xn} stands for the voltage reference corresponding to the x phase, and u_{c1}, u_{c2}, u_{c3} for the corresponding level-shifted carriers. Each PWM voltage reference corresponds to a min-max optimization strategy, which ensures improvement of the corresponding duty cycles, thus reaching the maximum possible modulation index by including the effect of common mode voltage, as presented in eq. (5)

$$u^*_{xn} = v^*_{xn} - \frac{1}{2}\left\{min(v^*_{an}, v^*_{bn}, v^*_{cn}) + max(v^*_{an}, v^*_{bn}, v^*_{cn})\right\} \tag{5}$$

As presented in eqs. (1) - (4) using the LS-PWM scheme only switching states 1, 2, 4, 8 are synthesized, neglecting the extra degrees of freedom that offers the converter topology with the redundant switching states, presented in Table 1. When low modulation index operation is required (for the proposed topology $m \leq 0.33$), because only the lowest voltage levels are being synthesized, a voltage unbalance in the dc-link starts to build-up, thus requiring the implementation of an additional dc-link voltage balance strategy, as reported in [9, 12, 13].

To deal with the previous described problem, the double star converter extra degrees of freedom can be exploited by implementing a PS-PWM modulation scheme, thus introducing a phase-shifting for the kth carrier denoted by θ_k and given in eq. (6)

$$\theta_k = (k-1)\frac{2\pi}{\ell-1} \quad k \in [1 \dots \ell-1] \tag{6}$$

where ℓ represents the number of voltage levels of the converter. All available switching states shown in Table 1 are then used. Redundant switching states are synthesized, according to eq. (5), (7) - (9), introducing redundant switching states to those allowed in the LS-PWM scheme.

$$S_1 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c1}(\theta_1)| \\ 0 & |u^*_{xn}| < |u_{c1}(\theta_1)| \end{cases} \tag{7}$$

$$S_2 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c2}(\theta_2)| \\ 0 & |u^*_{xn}| < |u_{c2}(\theta_2)| \end{cases} \tag{8}$$

$$S_3 = \begin{cases} 1 & |u^*_{xn}| \geq |u_{c3}(\theta_3)| \\ 0 & |u^*_{xn}| < |u_{c3}(\theta_3)| \end{cases} \tag{9}$$

PS-PWM scheme ensures that each capacitor will be delivering energy at its corresponding duty cycle, independently of the modulation index, thus reducing the natural voltage unbalance, without the need of an external voltage drift control scheme.

It also has to be noted, that for this particular application in a medium voltage AC-drive, using field oriented control scheme, the modulation strategy has to deal with the dc-link voltage unbalance, without the use of a voltage-drift control, thus the converter voltage references are set by the FOC control scheme, as a function of speed (within a certain load) and flux.

3 Induction machine Field Oriented Control drive

Squirrel cage induction machines (SCIM) are widely used in heavy industrial environments and applications. The three-phase SCIM mathematical model in the natural stator reference frame $\alpha\,\beta$ in state variables $\boldsymbol{i_s^{(\alpha\beta)}}$, $\boldsymbol{\psi_r^{(\alpha\beta)}}$, ω_r in the state space form (10)

$$\begin{aligned} \frac{d}{dt}\boldsymbol{x}(t) &= f_c\left(\boldsymbol{x(t)}, \boldsymbol{u(t)}\right) \\ \boldsymbol{y(t)} &= g_c\left(\boldsymbol{x(t)}, \boldsymbol{u(t)}\right) \end{aligned} \tag{10}$$

$$\boldsymbol{x(t)} = \begin{bmatrix} i_s^\alpha & i_s^\beta & \omega_r \end{bmatrix}^T \quad \boldsymbol{u(t)} = \begin{bmatrix} v_s^\alpha & v_s^\beta \end{bmatrix}^T$$

$$f_c = \begin{bmatrix} -\frac{1}{\tau_\sigma'} i_s^\alpha + \frac{L_m}{L_r r_\sigma \tau_\sigma' \tau_r} \psi_r^\alpha + \frac{L_m}{L_r r_\sigma \tau_\sigma' \tau_r} \omega_r \psi_r^\beta + \frac{1}{r_\sigma \tau_\sigma'} v_s^\alpha \\ -\frac{1}{\tau_\sigma'} i_s^\beta + \frac{L_m}{L_r r_\sigma \tau_\sigma' \tau_r} \psi_r^\beta - \frac{L_m}{L_r r_\sigma \tau_\sigma' \tau_r} \omega_r \psi_r^\alpha + \frac{1}{r_\sigma \tau_\sigma'} v_s^\beta \\ \frac{L_m}{\tau_r} i_s^\alpha - \frac{1}{\tau_r} \psi_r^\alpha + \omega_r \psi_r^\beta \\ \frac{L_m}{\tau_r} i_s^\beta - \frac{1}{\tau_r} \psi_r^\beta - \omega_r \psi_r^\alpha \\ \frac{3}{2}\frac{p}{J}\frac{L_m}{L_r} \psi_r^\alpha i_s^\beta - \frac{3}{2}\frac{p}{J}\frac{L_m}{L_r} \psi_r^\beta i_s^\alpha \end{bmatrix}$$

$$g_c = \begin{bmatrix} i_s^\alpha & i_s^\beta \end{bmatrix}^T$$

Control goals for the induction motor drive can be summarize as the following:

1. Maximum torque per Ampere operation.
2. Control of nominal flux.
3. Control of rotor speed.

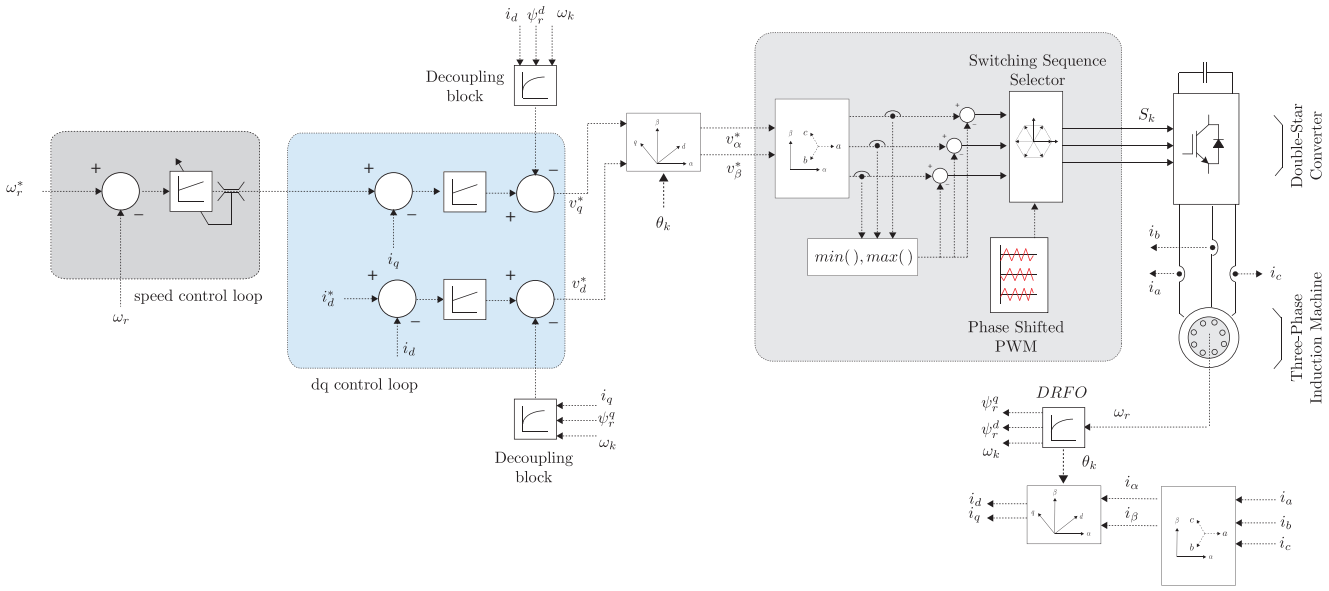


Figure 3: Field Oriented Control scheme with non-linearities compensation

The characteristics of FOC, have made this control strategy the most widely used for high demanding industrial applications [11]. Field Oriented Control (FOC) is based in the decoupling of the current space vector into a flux producing current and a torque producing current [14]. This is achieved by rotating the current space vector $i_s^{\alpha\beta}$ into a synchronous rotating reference frame dq, which is oriented by the rotor flux linkage space vector. Rotation of the state variables is achieved by means of the rotation matrix U given in eq. (11)

$$U = \begin{bmatrix} \cos\theta_k & \sin\theta_k \\ -\sin\theta_k & \cos\theta_k \end{bmatrix} \tag{11}$$

The dynamics in the dq synchronous reference frame are obtained by applying eq. (11) to eq. (10) . State variables $\boldsymbol{\psi}_r^{(dq)}$ $\boldsymbol{i}_s^{(dq)}$ can then be obtained as in (12), (13).

$$\boldsymbol{\psi}_r^{(dq)} = U \cdot \boldsymbol{\psi}_r^{(\alpha\beta)} \tag{12}$$

$$\boldsymbol{i}_s^{(dq)} = U \cdot \boldsymbol{i}_s^{(\alpha\beta)} \tag{13}$$

The implementation of the IM FOC control scheme is shown in Fig. 3. Compensation of non-linearities was achieved by including feed-forward compensation for the dq control loop PI controllers.

4 Simulation Results

Dc-link voltage stability performance simulation results, for the double-star converter, are presented for an industrial application consisting in a field oriented controlled induction machine industrial drive, under variable load and speed conditions.

Constant load and variable speed operation results are presented in Fig. 4. Simulations include the initial system speed start-up from 0 to 0.5 $[pu]$ of rated speed in ① and a speed step-up from 0.5 to 1.0 $[pu]$ of rated speed in ② under a linear torque characteristic such as $T_L = k\,\omega_r$

As shown in Fig. 4, during the speed start-up at ① using LS-PWM strategy the Upper and Middle capacitor voltages raise, because of the floating condition of ground, meanwhile the Lower capacitor voltage drops due the fact that only the lowest voltage level is being used and all energy is drawn only from this capacitor. On the other hand, when implementing PS-PWM scheme, there is no voltage drift. During the speed step-up at ② implementing LS-PWM the initial observed voltage drift increases, symmetrical from each other. With PS-PWM there is a natural voltage drop in all levels with a small voltage drift but only in the Lower capacitor.

Results for variable load and constant speed operation are shown in Fig. 5. They include the performance

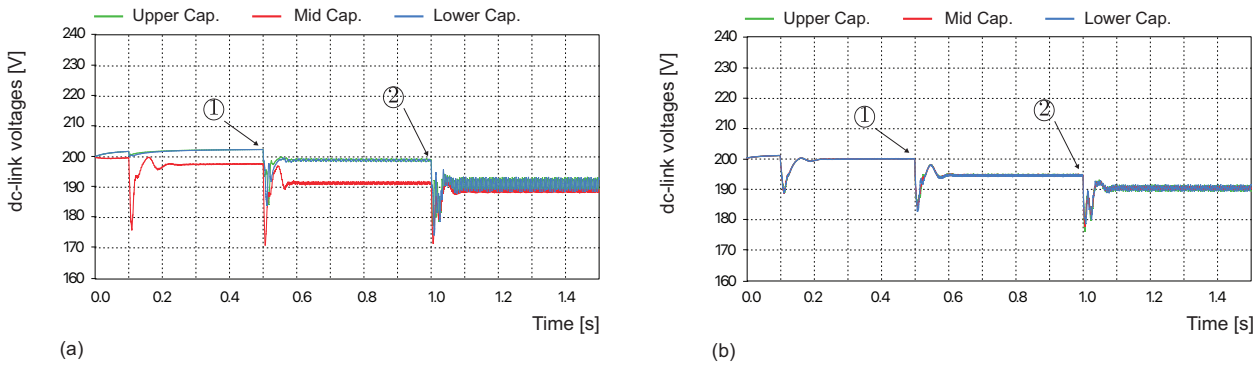


Figure 4: Simulation results for constant load-variable speed operation: (a) dc-link voltages using LS-PWM modulation (b) dc-link voltages using PS-PWM modulation

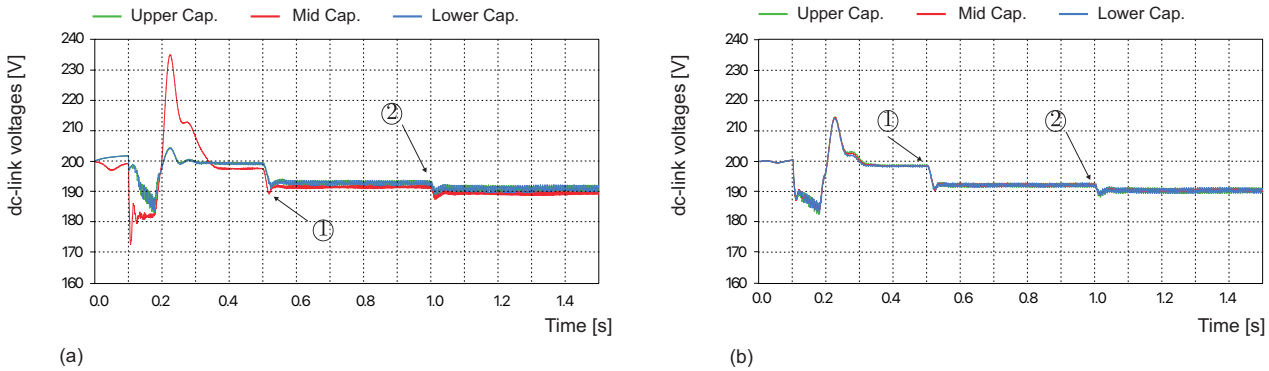


Figure 5: Simulation results for load impact under constant speed operation: (a) dc-link voltages using LS-PWM modulation (b) dc-link voltages using PS-PWM modulation

of dc-link capacitors voltages and the line - line output voltage, for the following operational conditions: initial speed-up transient under no-load condition, a first load impact of 0.6 [pu] at ① and a second load impact of 1.0 [pu] of rated load at instant ②

An evaluation of the actual benefits of PS-PWM strategy for this particular application in the DSC, can be made by comparing the capacitors dc-voltage drift and the output current THD, for both LS-PWM and PS-PWM strategies. These results are shown in Fig. 6 and in Fig. 7 respectively.

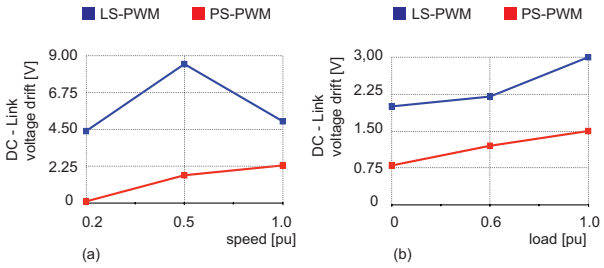


Figure 6: Total dc voltage drift: (a) constant load - variable speed (b) constant speed - variable load

The dc-link voltage drift is shown, for variable speed, constant load operation shown in Fig. 6 (a) and in constant speed, variable load operation as presented in Fig. 6 (b); in both cases, PS-PWM generates a lower voltage drift as with LS-PWM (it has to be noted that no voltage-drift control is implemented).

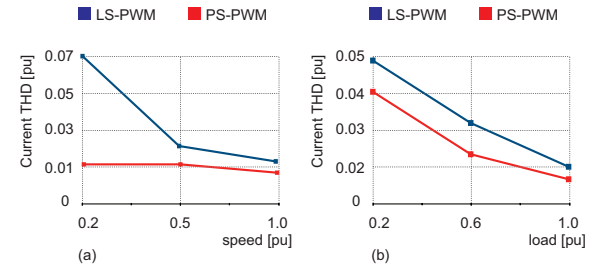


Figure 7: Line current Total Harmonic Distortion (THD), in per unit [pu] (a) constant load - variable speed (b) lconstant speed - variable load

In Figs.7 (a) and (b) the output currents THD performance is shown, for constant speed and constant load operation conditions, using both, LS-PWM and PS-PWM strategies. On each case the PS-PWM strategy presents a lower THD index.

The harmonic content of the output currents, has a directly impact on the electrical torque developed by the induction machine, due the fact that the electromechanical torque is a function of the stator currents and the rotor flux linkage space vectors.

In Fig. 8 the results for the i_d i_q current control loop, and the developed torque performance are presented. As shown in Fig. 8 (a) and (d) the i_d i_q current control loop (inner control loop) exhibits a good performance either using level shifted or phase shifted PWM. However, results for the electromechanical torque developed by the induction machine, when using LS-PWM strategy, shown in Fig. 8 (b) and (c), exhibit a more scattered harmonic pattern, with higher order frequency components, compared to the results obtained with the PS-PWM strategy, discussed previously.

As shown in Fig. 4 and Fig. 5 PS-PWM modulation strategy offers a better performance during the initial speed-up transition (low modulation index operation region), in terms of dc-link undershoot and voltage unbalance, when compared with LS-PWM modulation. Moreover, DC-link voltages show less voltage unbalance between each capacitor, resulting in lower THD for the output currents, as presented in Fig. 7.

Output currents with lower harmonic distortion, ensures low electrical torque ripple, as shown in Fig. 8 (c), so the torque harmonics are concentrated in the low frequency range, as presented in Fig. 8 (f), while, on the other hand, LS-PWM shows a very distributed torque harmonic profile, with high order frequencies, which can be a source of mechanical stresses.

5 Conclusions

In the present work a double-star converter application for a medium voltage induction motor drive under variable speed and load operation has been presented. The particular characteristics of the double-star topology, enables this multi-level voltage source converter to use the extra degrees of freedom, by synthesizing the redundant switching states via the phase-shifting PWM strategy.

The use of redundant switching states also improves the natural voltage-drift between the dc-link capacitors, due to its floating condition, resulting output currents with low harmonic content. Moreover, the developed torque shows a more symmetrical harmonic distribution, concentrated in the low frequency range, compensating the torque ripple, and thus preventing mechanical stresses on the output shaft. These characteristics make the double-star converter topology using PS-PWM modulation strategy, suitable for medium voltage drive applications.

The converter advantages for its application in medium voltage AC-drives, over classical medium-voltage voltage-source converters topologies, are based in the simplicity of its topology (no flying capacitors and clamping devices are required) and to fact that redundant switching states are available via conventional pulse-width modulation methods, reducing harmful high order torque harmonics.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The author also wish to thank the financial support from the Chilean Found for Human Resource Development (CONYCIT) through its Ph.D. scholarships (CONICYT/21130448).

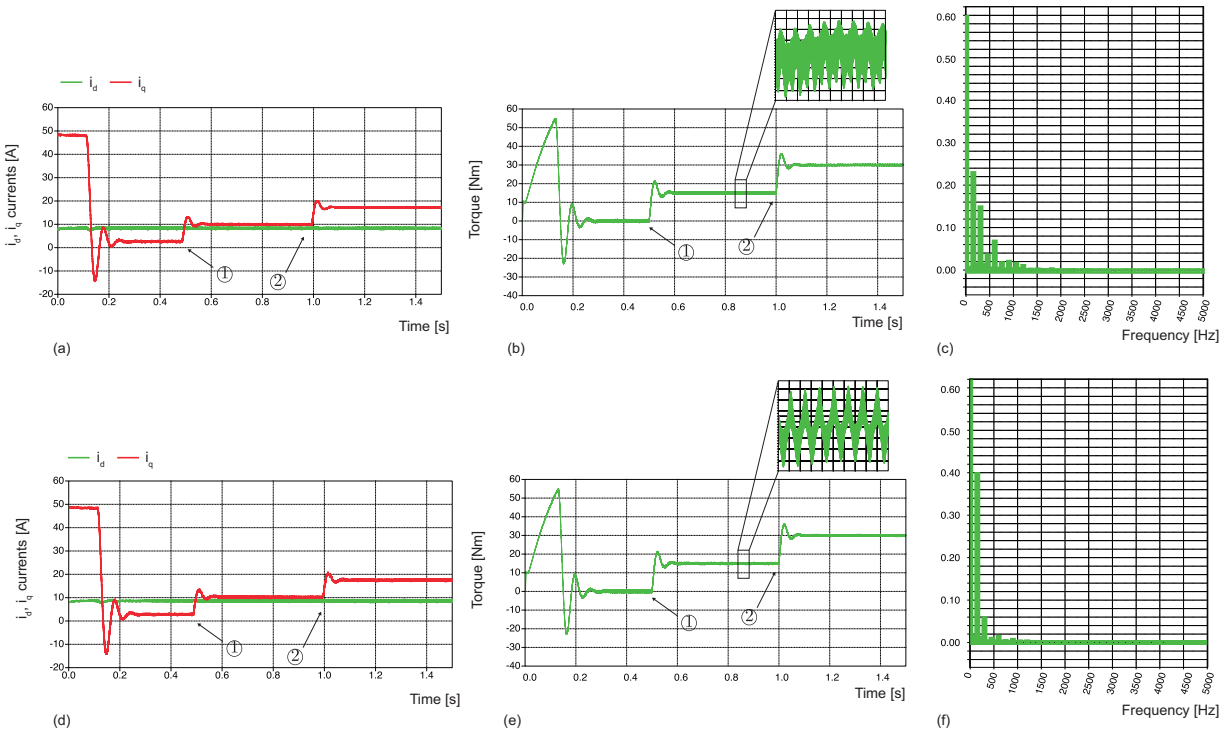


Figure 8: Simulation results under constant speed - variable load: (a) $i_d\, i_q$ currents using LS-PWM modulation (b) electric torque using LS-PWM modulation (c) torque harmonic profile using LS-PWM modulation (d) $i_d\, i_q$ currents using PS-PWM modulation (e) electric torque using PS-PWM modulation (f) torque harmonic profile using PS-PWM modulation

References

[1] I. Boldea and S. A. Nasar, "Electric drives," *CRC Press*, vol. 24, June 30 2016.

[2] J. I. Leon, S. Vazquez, and L. G. Franquelo, "Multilevel converters: Control and modulation techniques for their operation and industrial applications," *Proceedings of the IEEE*, vol. 105, no. 11, pp. 2066–2081, Nov 2017.

[3] S. Kouro, M. Malinowski, K. Gopakumar, J. Pou, L. G. Franquelo, B. Wu, J. Rodriguez, M. A. Perez, and J. I. Leon, "Recent advances and industrial applications of multilevel converters," *IEEE Transactions on Industrial Electronics*, vol. 57, no. 8, pp. 2553–2580, Aug 2010.

[4] L. G. Franquelo, J. I. Leon, and E. Dominguez, "New trends and topologies for high power industrial applications: The multilevel converters solution," in *2009 International Conference on Power Engineering, Energy and Electrical Drives*, March 2009, pp. 1–6.

[5] X. Yuan, "Derivation of voltage source multilevel converter topologies," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 2, pp. 966–976, Feb 2017.

[6] J. Rodriguez, S. Bernet, B. Wu, J. O. Pontt, and S. Kouro, "Multilevel voltage-source-converter topologies for industrial medium-voltage drives," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 6, pp. 2930–2945, Dec 2007.

[7] D. G. Holmes and T. A. Lipo, "Pulse width modulation for power converters: Principles and practice," *Wiley-IEEE Press*, vol. 24, October 3 2003.

[8] S. alberto Gonzalez, S. andres Verne, and M. I. Valla, "Multilevel converters for industrial applications," *CRC Press*, vol. 8, July 22 2013.

[9] S. Kouro, M. Perez, J. Rodriguez, and B. Wu, "Four-level medium voltage multilevel converter for high power applications," in *ECCE Asia Downunder (ECCE Asia), 2013 IEEE*, June 2013, pp. 1028–1033.

[10] L. M. Tolbert, F. Z. Peng, and T. G. Habetler, "Multilevel pwm methods at low modulation indices," *IEEE Transactions on Power Electronics*, vol. 15, no. 4, pp. 719–725, Jul 2000.

[11] W. Bin, "High-power converters and ac drives," *Wiley-IEEE Press; 1 edition*, pp. 1–468, 2006.

[12] C. A. Reusser, S. Kouro, A. Llor, R. Cardenas, and V. Yaramasu, "Permanent magnet synchronous generator wecs based on a four-level double star converter," in *2017 IEEE 8th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, April 2017, pp. 1–6.

[13] A. M. Llor, S. Kouro, C. Reusser, V. Yaramasu, and B. Wu, "Medium voltage 4-level double-star multilevel converter using model predictive control," in *2017 IEEE International Conference on Industrial Technology (ICIT)*, March 2017, pp. 136–140.

[14] J. Holtz, "The representation of ac machine dynamics by complex signal flow graphs," *IEEE Transactions on Industrial Electronics*, vol. 42, no. 3, pp. 263–271, Jun 1995.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 59-71 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

An FPGA Implementation of Resource-Optimized Dynamic Digital Beamformer for a Portable Ultrasound Imaging System

Jingwei Xu[*,1], Yi Zheng[2], Mohan Chirala[3], Mohamed Almekkawy[2]

[1]*Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA*

[2]*Department of Computer Science and Engineering, Penn State University, University Park, Pennsylvania, USA*

[3]*Samsung Research America, Dallas, Texas, USA*

ARTICLEINFO

Article history:
Received: 01 June, 2018
Accepted: 01 July, 2018
Online: 14 July, 2018

Keywords:
Medical Ultrasound
Digital Beamforming
Linear array
FIR Polyphase Filter
FPGA

ABSTRACT

This paper presents a resource-friendly dynamic digital beamformer for a portable ultrasound imaging system based on a single field-programmable gate array (FPGA). The core of the ultrasound imaging system is a 128-channel receive beamformer with fully dynamic focusing embedded in a single FPGA chip, which operates at a frequency of 40 MHz. The Rx beamformer is composed of a midend processing module, a backend processing module, and a control block. The midend processing module is established using the implementation of the delay summation through coarse delays and fine delays, with which delays could vary continuously to support dynamic beamforming. In order to enhance spatial and contrast resolution, the Rx beamformer is further accommodated by employing a polyphase filter, which improves the effective beamforming frequency to 240 MHz. The control block generates control signals based on a memory management block, which doubles the data transfer rate. The processed data is wirelessly sent to a commercial Android device. The low cost ultrasound imaging system supports real-time images with a frame rate of 40 fps, due to the limitation imposed by the wireless backhaul process. To reduce power consumption, a dynamic power management technique is used, with which the power consumption is reduced by 25%. This paper demonstrates the feasibility of the implementation of a high performance power-efficient dynamic beamformer in a single FPGA-based portable ultrasound system.

1 Introduction

Ultrasound imaging is the most commonly-used noninvasive real-time diagnostic tool in clinical applications because of its free of radiation and ease of use [1]. However, traditional ultrasound devices are bulky because of the large amount of digital data that needs to be simultaneously processed. For greater operational convenience, low power hand-held devices have come into prominence [2–4]. To minimize the size of ultrasound imaging systems, research laboratories continue to propose approaches based on modern signal and image processing methods. Their intent is to improve the image quality and diagnostic accuracy, without compromising imaging performance [5]. Medical ultrasound imaging has been used for imaging organs in human body and measuring the blood flow in vessels through Doppler shift [6].

A number of approaches have been reported for ultrasound imaging using both linear array and phased array [7–11]. One-dimensional ultrasound transducer arrays are used to get 2D images, and two-dimensional transducer arrays have been applied to produce 3D ultrasound images [12–14]. The ultrasound produces images through the backscattering of the mechanical energy from boundaries through tissues. Because of this property, a 1D image could be produced by 1 transducer element exciting ultrasound waves along a

[*]Corresponding Author: Jingwei Xu, Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas USA, Email: xujw07@tamu.edu

straight line, called scanline (SL), and a 2D image could be generated by repeating the same step for all transducer elements. With a frame rate of approximately 30 fps, ultrasound imaging is considered to be a real-time imaging technique. However, the relatively poor soft tissue contrast limits the performance of ultrasound imaging technique. In addition, the increase in integration density tends to increase the power consumption, which is a major constraint in the implementation of digital signal processing (DSP) architectures especially for hand-held ultrasound imaging system. Although some ultrasound devices are implemented based on application-specific integrated circuits (ASICs) to reduce cost and power consumption [15], the ASIC-based design is not preferred by researchers due to its inflexibility for advanced applications. On the contrary, FPGA, because of its reconfigurability, has become an ideal alternative to ASIC. There are some existing FPGA based designs on ultrasound, but these designs are limited by several criteria discussed below.

First, it is hard to implement the DSP portion of the system with a low hardware complexity, while not impacting the spatial and temporal resolutions of the system. In fact, portable ultrasound imaging systems are limited by the low resolution and high power consumption. Increased number of transducer elements, though providing higher resolution, making the entire data rate at ADC interface reach a forbidding level that challenges not only the power consumption but also the data transmission ability. In addition, such large transducer array requires hugh analog-front-end (AFE) and digital circuitry operates in parallel, such as low-noise-amplifier (LNA), time-gain controller (TGC), and analog-to-digital converter (ADC). However, the large number of hardware duplication may cause a lot of implementation issues when integrated into portable form factors, including heat dissipation, crosstalk interface, I/O packaging. Several prior methods have been developed to reduce hardware complexity. A 64 channel power-efficient architectures version without the real time controller (RTC) has been shown by [16]. Other schemes described the DSP algorithms used in digital beamforming [17–19]. In addition, an ASIC-based programmable ultrasound image processor is shown by [20]. A 16-channel FPGA-based real-time high frequency digital ultrasound beamformer is discussed by [21], but it could only support 30 fps frame rate, which constrains the image quality.

Seconds, the temporal resolution is bounded by the bandwidth of the ultrasound imaging system, which is limited by the ADC implementation [22, 23]. The temporal resolution in digital beamforming systems can typically be improved in two ways: either using phase rotation, also known as direct sampled in-phase/quadrature (DSIQ) beamforming, or through interpolation-based filtering. However, phase rotation degrades image quality as it assumes the ultrasound signal to be a narrow band [4, 24], and interpolation-based filtering requires up-sampling before low-pass filtering, which requires the operation of the DSP portion on high frequencies, and therefore raises the energy cost. Therefore, we need to improve the methods above to reduce power consumption.

Third, applying exact delays on corresponding signals is a challenge, because the amount of delay between two subsequent sampled signals may vary based on the nature of dynamic beamforming. An iterative algorithm for calculating delay information in real-time has been implemented by [25]. However, the delay resolution in that system is bounded by the frequency of the delay calculation module. Therefore, a higher operating frequency is necessary to acquire more accurate delay information. Another widely used method is precomputing the delays. However, to overcome the impending memory limitation of FPGAs, only a pseudo-dynamic focusing technique is utilized where delay information is only updated for a pre-determined depth rather than for a pre-determined delay resolution [26]. Different articles introduced the digital beamformer design using a 5 MHz center frequency beamformer linear array implemented on FPGA as discussed by [27] and with a 50 MHz center frequency annular array by [28,29]. A research-aimed FPGA-based digital transmit beamformer system for generating simultaneous arbitrary waveforms has been developed by [30]. Others have applied over sampling techniques and single transmit focusing as other methodologies for digital beamforming [31].

Therefore, there is a need to develop resource-friendly, power-efficient, and space-efficient architectures and methods for both properly delaying sampled signals and correctly acquiring delay information for a portable ultrasound device, where innovative designs of AFE and mixed-signal interface are keys to the next generation portable ultrasound devices. State-of-the-art portable ultrasound typically operates at 2 to 4 MHz with less than 64 active transducer elements.

In this paper, we present our 128-channel portable ultrasound imaging system embedded in an Arria V FPGA (Altera Inc., San Jose, CA, USA), where several methods are introduced, such as polyphase filter, memory management block, and active aperture, with which we resolve the issues discussed above so as to deliver a resource friendly ultrasound imaging system. Our ultrasound imaging system, compared to prevailing 32-channel [26] and 64-channel ultrasound imaging system [32], outputs higher resolution images with the better use of FPGA area. An overview architecture block diagram of the developed N-channel power-efficient and space-efficient ultrasound imaging system is shown in Figure 1. For the AFE part, a serial low voltage differential signaling (LVDS) interface protocol is used to get a fast analog acquisition, which serves as a serial way to recover parallel data. In addition, the echo-backed signal on each channel is amplified by a TGC amplifier to compensate for signal attenuation through propagation in the medium. The TGC amplifier is implemented on the AFE part as shown in Figure 2. The AFE implementation is out of the scope of this paper and will not be discussed in detail. The delay-and-sum (DAS) module in the midend processing module receives the compressed

timing information and outputs beamformed signal without decompression.

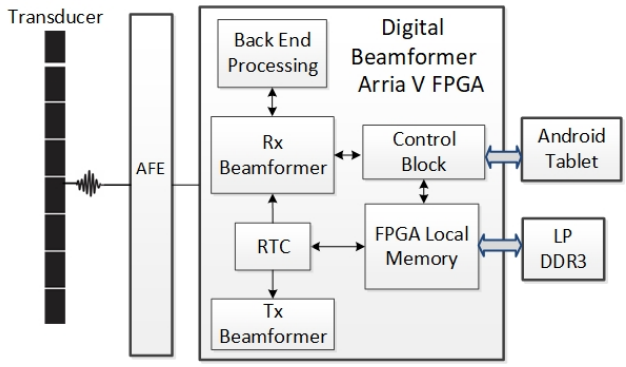


Figure 1: Architecture block diagram of the ultrasound digital beamformer based on a single FPGA.

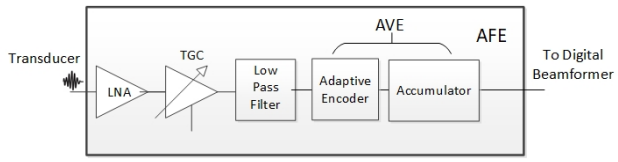


Figure 2: AFE architecture block diagram in the ultrasound digital beamforming system.

2 System Architecture

Figure 3 shows the detail of the beamforming architecture of the N-channel FPGA-based ultrasound imaging system. The midend beamformer will first pass the signal to S2P module to parallel the signal, since the AFE will serialize 2 signals from neighboring channels. Then, each signal will be passed into a coarse delay module, which delays the signal from the corresponding channel by integer numbers of the sampling period, T_s. The signal is then passed into a fine delay module, which delays the signal from coarse delay module by a fraction of T_s, and an apodization weight module, which multiplies the signal by a weight window function. The coarse delay block is achieved through controlling the writing process to the first-in-first-out (FIFO) memory to get valid data. The fine delay block is employed to replace the traditional way of optimizing the resource and is accomplished by a polyphase FIR filter. Finally, the processed signals from the N channels will be summed and sent to the backend processing module. The back end processing module processes acquired data based on an user-defined imaging mode, which can be selected via the Android tablet. The FPGA local memory is connected with an external memory, LP DDR3 to store the large amount of pre-computed delay information. The architecture of the ultrasound imaging system supports apodization and dynamic focusing, and is able to achieve a high-performance imaging system. In the following subsections, we will present the architecture of the Tx and Rx beamformer, the implementation of the midend processing module, the backend processing module, the control signal generation, and the calculation of the propagation delay.

2.1 AFE Architecture

AFE contains an LNA, an LVDS interface protocol, a TGC amplifier, and a low-pass filter, as shown in Figure 2. The LNA is used for amplifying small echoed back signals; LVDS interface protocol is used for getting a fast analog acquisition; the low-pass filter is used for cutting off the high-frequency noises. Inside the AVE, an adaptive encoder and an accumulator is employed. The accumulator detects transition edge of each piecewise-constant section and calculate the length. The adaptive encoder module performs adaptive thresholding and converts the amplitude variation into ternary timing information. The adaptive encoder consists of a comparator and a threshold generator. Two 1-bit ADCs are with in the comparator, and each receives input of the input signal and a low-threshold or a high-threshold signal, generated by the threshold generator. The comparator will outputs +1 or -1 if the input signal is greater than the high threshold or low than the low-threshold, respectively, and then requires the threshold generator to update the comparison parameter for the next comparison based on the input variation; otherwise, the comparator simply outputs 0. In this way, unit amplitude +1 or -1 could be assigned without loss of generality.

Some modulation methods, including the time encoding machine (TEM) [33], delta modulation [34], and integrate-and-fire scheme [35], converts amplitude information into timing information, and both TEM and delta modulation include a negative feedback loop to lock the input signal. In addition, these techniques will force the entire circuit keeps on flip-flopping, similar to a sigma-delta modulator output with constant input, and firing all the time even when there is no input signal. Also, typical integrate-and-fire scheme calculates the running average and also fires even when no variation occurs, resulting in unnecessary power overhead. Whereas in our adaptive encoder, the circuits fire only when significant variation occurs. As a result, the proposed adaptive encoder modulates the amplitude variations into delay information more efficiently.

2.2 Tx and Rx Architecture

The Tx beamformer supports 128 channel ultrasound beamforming signals, formed by two essential parameters: pulse-shape, assumed to be the same for all channels, and channel-delay, the value of which may vary among channels, and some channels could even be disabled. The two parameters are inputs to SPI ports, governed by the SPI clock. The SPI clock may run at different frequency compared to Tx clock which provides the Tx delay granularity.

Figure 4 shows the architecture of Rx beamformer, which receives control signals from both RTC and local memory management block, which will be discussed in

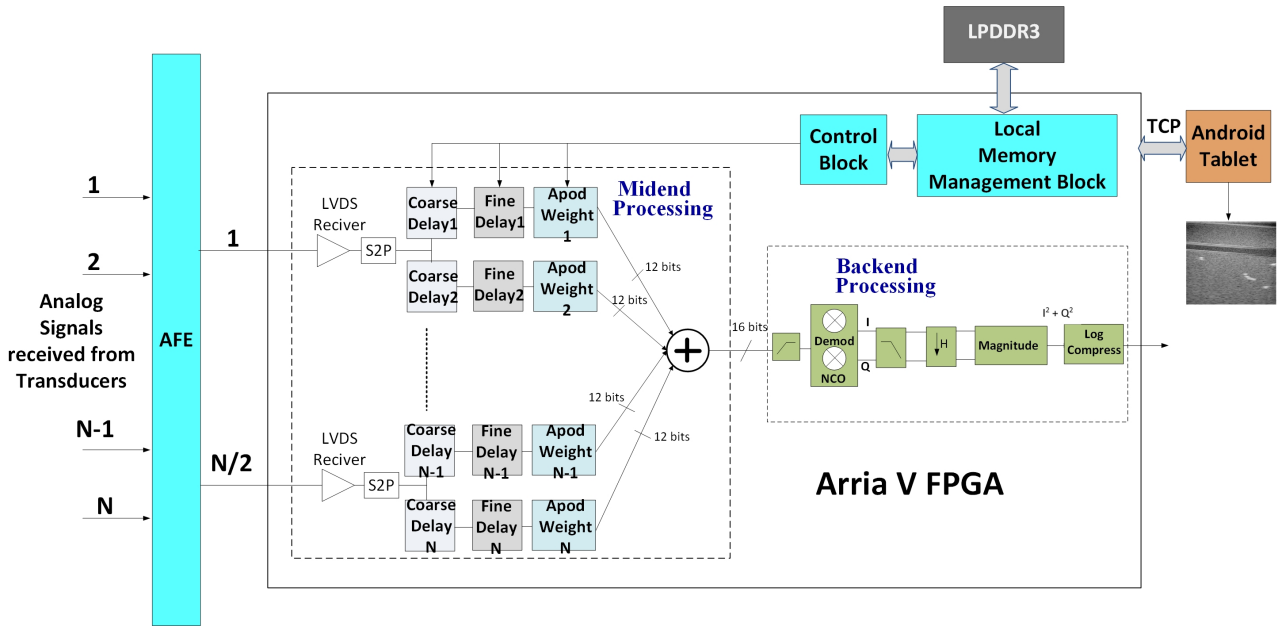


Figure 3: Architecture of ultrasound receiver of the single FPGA-based the ultrasound digital beamformer, showing the midend processing and the backend processing module with the control block and the local memory management block.

section 2.5, and outputs control signals to AFE. Eight sub-blocks made up the frame of the Rx block, and each comprises an LVDS receiver, a Frame Align module, a data management block, and a FIFO memory block. The Rx AFE SPI module outputs control signals serially.

2.3 Midend Processing Architecture

The core of midend processing module is coarse delay filters and find delay filters. The coarse delay filter delays the signals by the integer number multiplication of the sampling period, T_s. This block is implemented as a FIFO buffer. The coarse delay outputs a buffered signal when the desired control delay equals to the number of signals stored in the FIFO buffer. In the dynamic beamforming, some signals are used twice for the alignment. It is satisfied by this mechanism. As a result, if the control signal is greater or equal to the number of signal in the FIFO buffer, it is possible that when the control signal increment by 1, the coarse delay output remains the same. Hence, to get the same sampled signal, the coarse delay should be incremented by one as well, which prevents updating output signals. Therefore, by focusing the points from the nearest to the farthest along a scanline, the desired delays for all channels increase simultaneously.

Fine delay filter delays signals by a fraction of the sampling period, T_s. Figure 5 shows the architecture of the fine delay filter. As shown in Figure 5, the fine delay filter uses 48 FIR low-pass filters, from C0 to C47. Their parameters are calculated based on the baseband width and the carrier frequency. The 48 filters are grouped into 8 taps, with 6 phase coefficients per tap. Each set serves to multiply the input signals, or buffered signals, by the corresponding coefficients. Then, the multiplied signals are summed up to produce the interpolation signals.

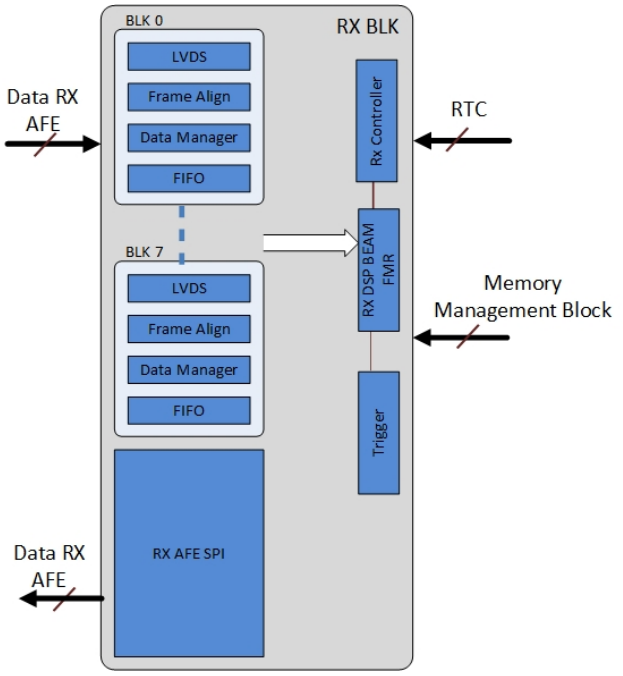


Figure 4: The Rx block diagram, showing Rx block is controlled by the control signal and RTC, and contains several subblocks 8 LVDS blocks, 8 frame align blocks, 8 data management blocks, and 8 FIFO memory blocks.

Based on the desired phase shift, the control signals generated by the control block select which set of coefficients will be used to multiply the input signals. The purpose of using register buffer is to reject the last repeated signals, because it is possible that when the coarse delay increases by 1 the output of coarse delay

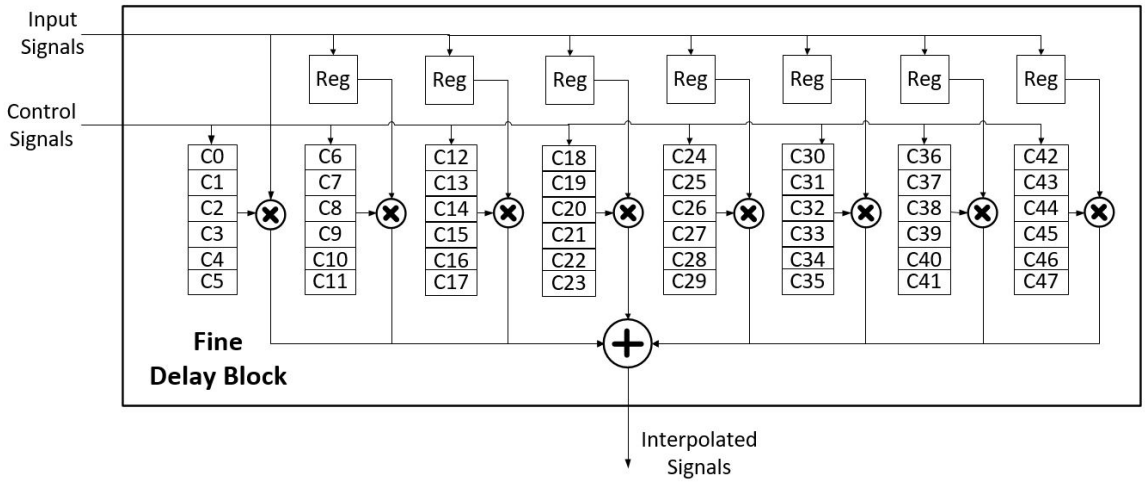


Figure 5: Block diagram of poly-phase fine delay filter with 7 registers for rejecting the last repeat signals and 8 × 6 low-pass filters for interpolation.

blocks are not updated. Therefore, the clock-gating scheme is introduced to prevent such hazard. Once the coarse delay is increased by 1, the pipeline enable signal will be pulled down to gate the clock. Therefore, the pipeline in the fine delay block will be stalled in one cycle to reject data input.

2.4 Backend Processing Architecture

Figure 3 also shows the architecture of the backend processing module. The backend processing module includesa high-pass filter, a demodulator, a numerically controlled oscillator (NCO), and a low-pass filter. The parameters of demodulation block are determined using Matlab (TheMathwork's Inc., Natwick, MA, USA). The data in this block is beamformed in 2 steps: before quadrature demodulation and after quadrature demodulation. The first step is the high pass filtering followed by the low pass filtering. This allows users to define filter option through the Android GUI. The second step architecture contains a down-sampling block, a magnitude I/Q demodulation block, and a log compression block. The I/Q demodulation block is used to extract the amplitude from the compressed data and outputs the interpolated data, where I and Q are 12 bits signed signal. The magnitude calculator outputs $I^2 + Q^2$, a 23-bit value. The log compression block is used to compress the data by taking its natural log value, which reduces the dynamic range, and outputs a 8-bit value.

2.5 Control Signals Generation

Control signals are generated by 2 blocks: RTC and the control block. Figure 6 is an overview of RTC. RTC contains 2 major blocks: the memory and RTC finite state machine (FSM). RTC FSM generate control signals based on data loaded from memory and controls Tx and Rx block along with memory.

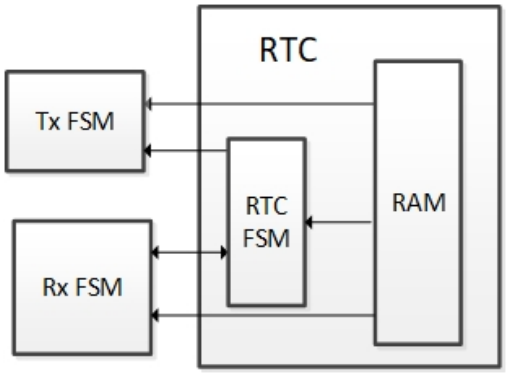


Figure 6: RTC block diagram showing the how RTC communicates with Tx and Rx block.

All N datapathes in Figure 3 are controlled by the control block. The control block generates control signals based on information passed from the local memory management block, where the FPGA local memory locates. An external memory is used to accommodate the FPGA local memory to store the large amount of pre-computed delay information. The corresponding delay information will be loaded into FPGA local memory based on the imaging mode that is selected on the user side, and the command will be sent wirelessly to the control block. The control block controls all delay filters by loading pre-computed data stored in the FPGA local memory. The FPGA local memory updates itself by reading data from an external memory. However, since it takes much more time to retrieve data from the external memory than from the local memory, it is necessary to adjust the data load process, necessitating the memory management block. Figure 7 shows the architecture of the memory management block. The FPGA local memory receives data from the external memory. The input data is passed into a demux, which is controlled by the number of scanlines processed. Each demux output is connected to half of the FPGA local memory, the output of which is connected to a mux, which is controlled by the num-

ber of scanlines processed. The output of the mux is the output of this block. Local memory A and B are used as buffers, where A is used to buffer the current scanline and B is used to buffer the next scanline. The demux will pass input data to local memory A if the number of scanline processed is an odd number, and to B otherwise, represented by $2k+1$ and $2k$, respectively, where k is a non-negative integer. The mux will output data from local memory A if the number of scanline processed is an even number, and from local memory B otherwise. With this technique, the FPGA local memory is always passing the current scanline data and next scanline data, reducing the time spent on transferring data from the external memory to the control block.

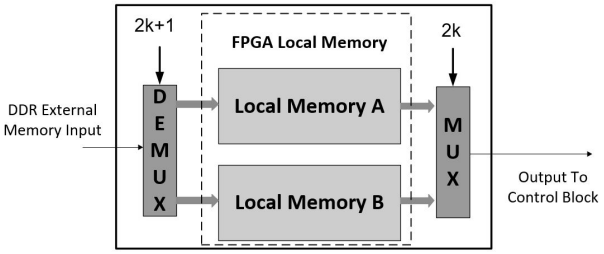


Figure 7: Block diagram of local memory management system for control signal generation process.

3 Propagation Delay Calculation

Beamforming is a algorithm which takes inputs from the transducer elements and outputs the summed signals. The implementation of beamforming algorithms is one of the deterministic factors in the ultrasound imaging quality [36]. In this paper, a delay-and-sum beamformer is used, because of its high accuracy on close-range beamforming, which is a dominant in medical ultrasound systems. Several focusing techniques have been developed for delay-and-sum beamforming, including the fixed focusing, dynamic focusing, and composite focusing techniques. The composite focusing technique is simply a special case of dynamic focusing [37] and will not be discussed in this paper.

Figure 8 shows how the signals alignment system (SAS) works. As Figure 8 presents, the signals echoed back from the focal point reach different channels at different time, which causes delay. The SAS module is used to align the signals and then pass them into the summer, which outputs the summed signal.

The difference between dynamic focusing and fixed focusing is that dynamic focusing requires the focal point to be continuously changed. Dynamic focusing could extend the field depth without satisfying the frame rate [38]. Therefore, the dynamic focusing technique is used when the SAS operates in the receive phase, and the delays of all channels are adjusted, because of the requirement of dynamic focusing. Figure 9 demonstrates the changes in the signals' flight paths during the receive phase. In Figure 9, multiple channels are used to receive ultrasound signals. Focal points on the same scanline are under detection. By assuming the scanline begins from the position of channel A, we can determine the focal points of the signals sampled by channel A. There exists a depth in the physical positions between every two sampled data points which is clearly evident in Figure 9. The distance, d, between two focal points, P_1 and P_2, satisfies that $d \approx (c \times T_s)/2$, given that d is small enough. Here, c is the average ultrasound speed in human body.

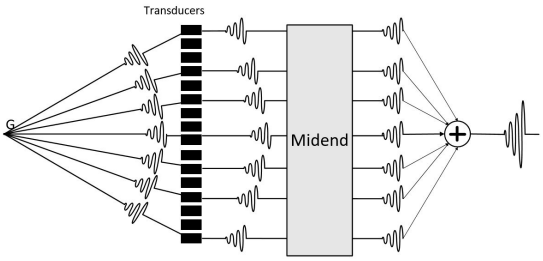


Figure 8: Signals alignment in ultrasound digital beamforming system.

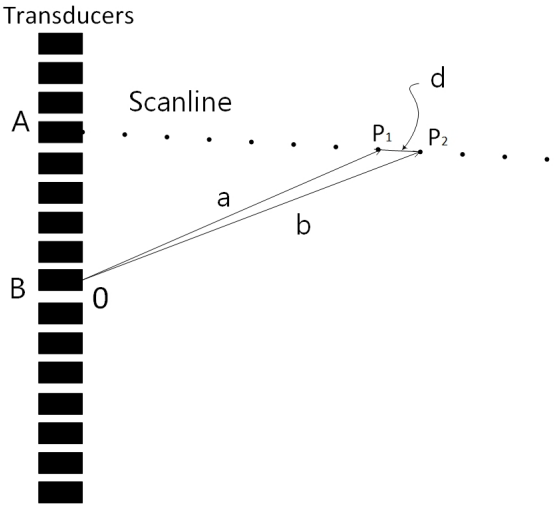


Figure 9: Variations of echo paths in receiving signals used in dynamic focusing technique.

Now, the signals that are variably received by the transducer elements should be delayed using dynamic beamforming. For the focal point P_1 in Figure 9, it is assumed that the time interval between transmission and reception by channel A is T_1. The time interval for channel B to receive it is T_2, where $T_2 > T_1$. Consequently, a delay of $(T_2 - T_1)$ in the signal echoed back from P_1 is observed to be aligned with the signal in channel B. For the subsequent focal point P_2, the time that corresponding signals are received by Channel A and Channel B, $\hat{T_1}, \hat{T_2}$ respectively, are:

$$\begin{aligned} \hat{T_1} &= \Delta T_1 + T_1 \\ \hat{T_2} &= \Delta T_2 + T_2 \end{aligned} \tag{1}$$

According to the property of the triangle, we have

$b - a < d$, where a is the distance between P_1 and channel B, and b is the distance between P_2 and channel B. Thus it is derived that:

$$\begin{aligned} \Delta T_2 &= \frac{d+b-a}{c} < \frac{2d}{c} \\ \Delta T_1 &= \frac{2d}{c} \end{aligned} \tag{2}$$

It implies that the corresponding signal of P_2 has arrived at channel B before its next sampling. We can retrieve the fraction delay between two samples after sampling the next signals in channel B provided that $\Delta T_1 - \Delta T_2$ is relatively short. But if we get a large enough $\Delta T_1 - \Delta T_2$, the corresponding signal has to be accessed without sampling the next signal. Here, we use the previous sampled signal from channel B twice and align with both signals received from the channel for P_1 and P_2. The resolution our system can support is $\lambda/16$. There are techniques that use a 100 MHz sampling rate with digital interpolation filters [21], which allow the signals to be sampled at a lower frequency and concurrently get more precise delays [39, 40].

The delay values are pre-computed and pre-loaded to the external memory, as mentioned in the previous section.

4 System Specifications

The T_x firing frequency, also called pulse repetition interval (PRI), is constrained by several theoretical limitations: the round trip delay (RTD), the wireless backhaul limitation (WBL), and the DDR to FPGA interface limitation.

First, PRI is given by:

$$PRI = \frac{1}{N_s \times FR} \tag{3}$$

where N_s refers to the number of samples per scanline, and FR is the frame rate of the system.

N_s could be calculated by:

$$N_s = f_s \times RTD \tag{4}$$

where f_s indicates the ADC sampling rate, as defined previously, and RTD is calculated by 5,

$$RTD = \frac{2l}{c} \tag{5}$$

where l is the penetration depth. The penetration depth varies among different mediums. Theoretical time for penetration depth is shown in Table 1:

WBL is given by

$$WBL = \frac{N_s N_b}{W_b} \tag{6}$$

where N_b is the number of bits per sample; and W_b, the wireless bachhaul speed, is determined by the device employed. The WBL indicates the lower bound for PRI.

Table 1: Theoretical RTD for Different Penetration Depth.

Depth of Penetration (cm)	RTD (μs)
9	116.88
7	90.10
5	64.94
3	38.96

5 Dynamic Power Management

To reduce power consumption, we modified the traditional B-mode imaging to perform the following scans: for each focal point on a specific scanline, excite a pulse from the transducer corresponding to that scanline, and receive signals with all transducers within active aperture. Repeat this procedure for all scanlines. The reason this method is able to reduce power consumption and conserves the image quality is because of 7:

$$F_n = \frac{f}{D} \tag{7}$$

where F_n is the f-number, a fixed number for one transducer, f is the focal depth and D is the diameter of the hypothesis aperture. According to 7, the maximum hypothesis aperture size is proportional to the depth of the furthest focal point, which indicates that channels outside the hypothesis aperture does not help the preservation of image quality. Figure 10 shows how this equation helps reduce the power consumption. As shown in Figure 10, for focal point B, the hypothesis aperture is the black region in the center. When focal point B is processed, all transducers outside the hypothesis aperture could be turned off. When the focal point is changed from B to A, the size of the hypothesis aperture increases, so more channels could be turned on to receive signals. The transducers in the white region are still off, because they are outside the hypothesis aperture for focal point A.

To further reduce power consumption, we consider turning off additional transducer elements with negligible image quality degradation. The idea is to choose a range of transducer elements, the range of which is called limited aperture, shown in Figure 11, and listen from the transducer elements within the limited aperture. However, there are cases where the transducers are not enough to cover the whole limited aperture, as illustrated in Figure 12. In this case, the aperture covering both transducers and limited aperture is called active aperture, and only the channels within the active aperture are turned on.

To get better image quality, the received signal for a focal point is the sum of all signals received by different active transducer elements. This can be achieved by a simple transducer elements shift operation shown in Figure 13. When processing SL_1, the transducer elements within the active aperture only are turned; for SL_2, the limited aperture is shifted down by 1 transducer element, so that one more transducer element

is turned on to receive signals; and when SL_n is processed, all transducer elements within the limited aperture are turned on. This shift operation will last until all scanlines are processed.

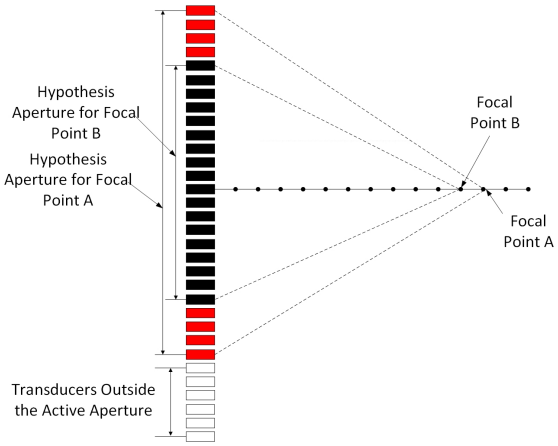


Figure 10: Hypothesis aperture for different focal points on one scanline, showing the variation of hypothesis aperture with the change of focal depth.

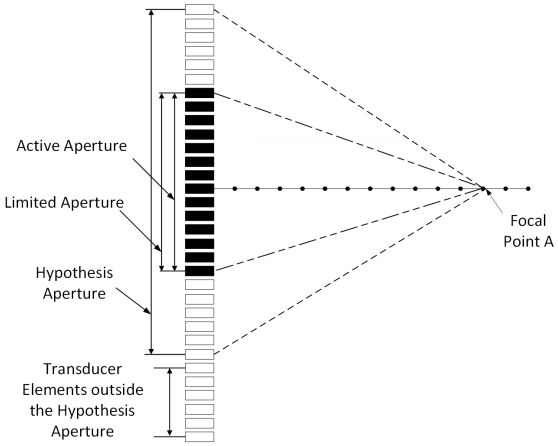


Figure 11: Limited aperture size for further focal points showing the relationship between active aperture, limited aperture, and hypothesis aperture in the normal case.

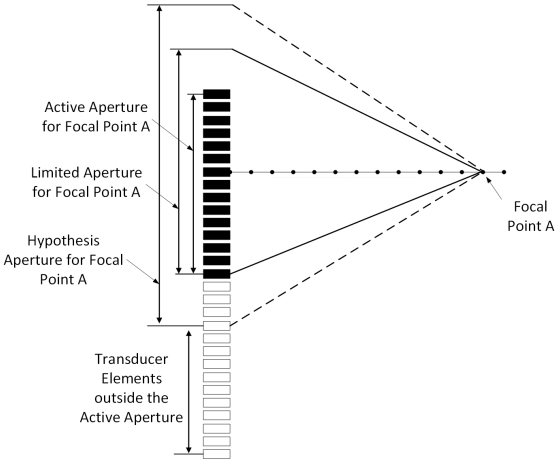


Figure 12: Limited aperture size for further focal points showing the relationship between active aperture, limited aperture, and hypothesis aperture when the transducer elements cannot cover the whole limited aperture.

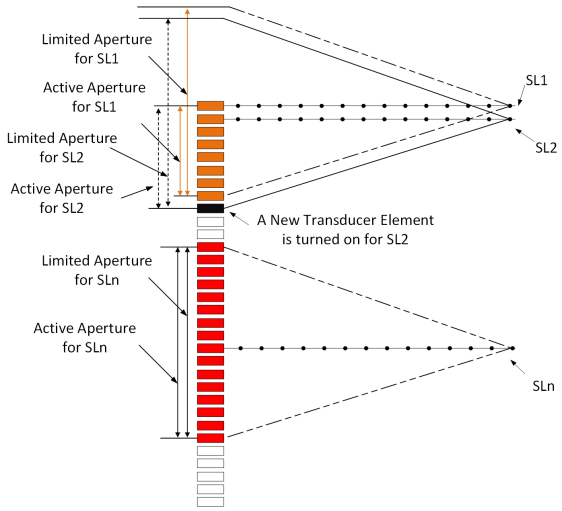


Figure 13: Transducer elements shift operation illustration. The limited aperture will be shifted based on the location of scanlines.

For a 128-channel ultrasound system, the maximum hypothesis aperture size could reach up to 256, including a virtual array consisting of 128 elements (the intersection between hypothesis aperture and aperture out of limited aperture) and physical aperture consisting of 128 elements. To observe the degradation on performance, different sizes of the limited aperture are used to receive echoed back signals. The corresponding images are shown in Figure 14. These images show that with the receive aperture increasing, a higher imaging contrast resolution could be achieved. When the limited aperture is larger than half of the effective aperture (the whole transducer array), the degradation is conceptually negligible. With the aperture size less than half of the effective aperture, most parts of the image such as blood vessels could still be clearly imaged. However, the bright spots are spread around. Hence, when the aperture size increases, a better signal-to-noise ratio (SNR) is achieved. To acquire a good measurement of the degradation, the formula below is applied to quantitatively check the difference among those images.

$$CR = 10log_{10}[\frac{\bar{I}_2}{\bar{I}_1}] \tag{8}$$

where CR is the contrast ratio, and $\bar{I}_1, \bar{I}_2$ are the average intensities in the region of interest and the region of reference, respectively. This provides a quantitative assessment of the performance of contrast restoration, the result of which is shown in Table 2.

As we discussed above, we know that the image quality will vary depending on the active aperture size. Thus, for the case we illustrated, we could merely keep the 128-channel aperture instead of the 256-channel aperture. Therefore, based on the analysis above, we could turn off all transducers outside the active aperture to reduce power consumption, while still conserving the image quality.

To estimate the power consumption, we assume that each transducer element could be independently

Table 2: Contrast Ratio of the Images with Different Limited Aperture Values.

Aperture	128	96	64	32	16	8
CR	16.3842	16.0634	15.4241	14.7033	14.637	14.637

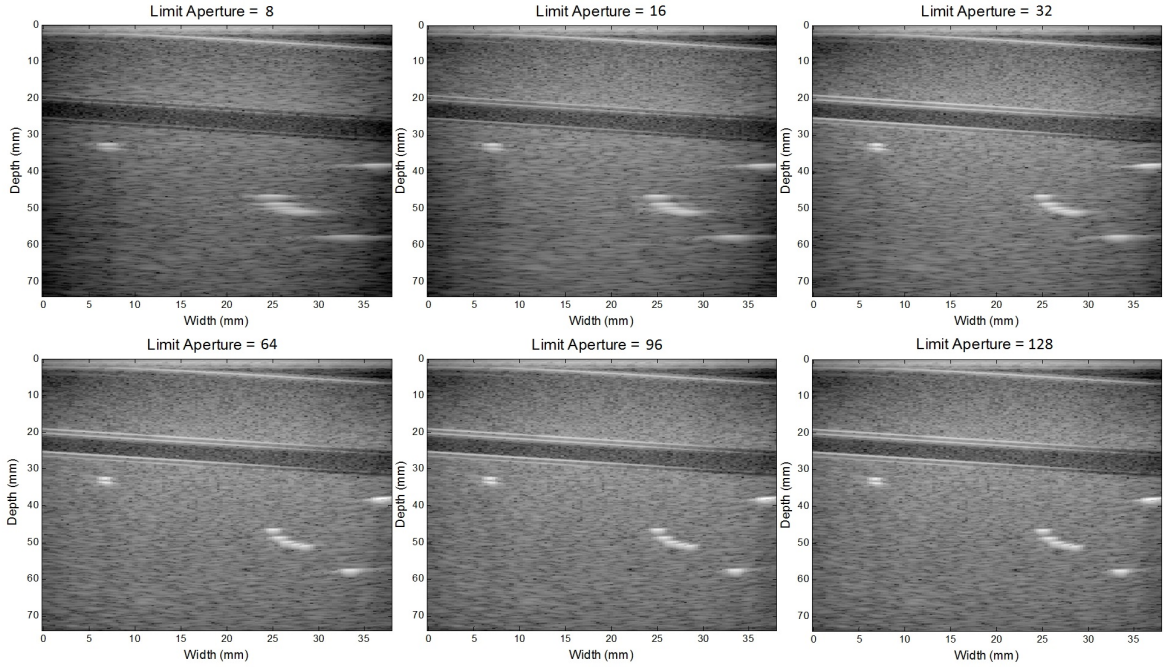


Figure 14: Ultrasound images generated with different limited aperture size chosen, showing that the image quality reserves when the limited aperture size is greater than or equal to half of the effective aperture.

switched "on" and "off" by using power-gating technology. For a linear array, the power consumption is approximately proportional to the total number of active elements used during scanning one frame:

$$Power \propto \sum_{i=1}^{128} N_i \tag{9}$$

where N_i is the number of data channels used for the ith scanline. N_i could be represented by

$$N_i = Ch_{right} - Ch_{left} + 1 \tag{10}$$

where Ch_{right} and Ch_{left} are the indexes of right most and left most active transducer element, respectively. They can be represented as in 11 and 12.

$$Ch_{right} = \begin{cases} i + LA & \text{if } i + LA < 128 \\ 128 & \text{if } i + LA \geq 128 \end{cases} \tag{11}$$

$$Ch_{left} = \begin{cases} i - LA & \text{if } i - LA \geq 1 \\ 1 & \text{if } i + LA < 1 \end{cases} \tag{12}$$

where LA is the size of limited aperture. The index of the very right channel is on the right of the channel which is in the position of the current scanline. The limitation of aperture allows the channels whose index is larger than $i + LA/2$ to be turned off. If the very right channel outside the zone of the transducer, i.e. whose index is larger than 128. The 128th channel will be the right bound.

According to 9, the normalized power consumption for different aperture limitations are calculated and plotted in Figure 15. With negligible degradation, the 128-channel ultrasound device will consume 75% of the power that the all-channel effort devices consume.

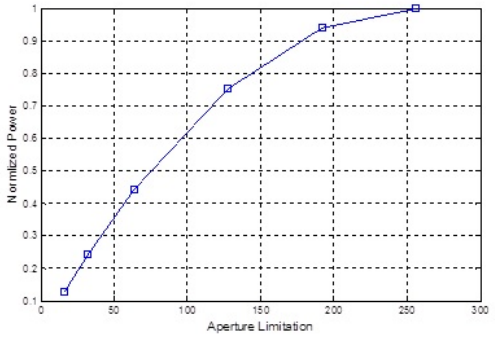


Figure 15: Normalized power consumption for different limitations of aperture.

6 Results and Discussion

The timing of the beamforming architecture is shown in Figure 16. Initially, the system is powered up, allowing switching mode power supplies (SMPSs), AFE ICs, FPGA, and wireless to be on. When the user selects the smart probe function via tablet GUI, a data packet containing initial system configuration is sent to the RTC wirelessly. The RTC will extract the imaging mode, application, array type, FR, Ns, and number of scanlines (NSL) from this packet. Then the RTC will continue to

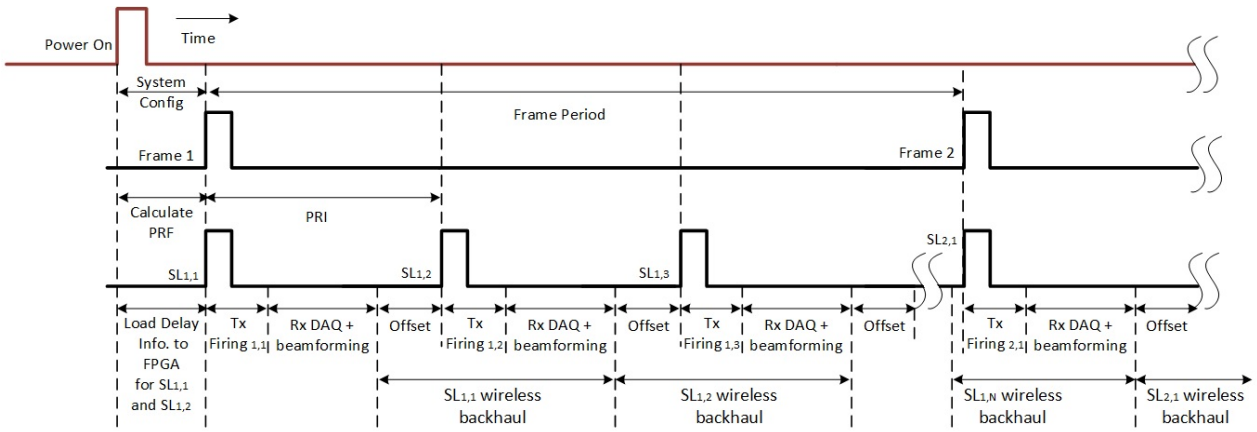


Figure 16: PRI adjustability based on system requirements, showing the conceptual timing of the single-FPGA-based ultrasound imaging system.

load and send delay information for SL_1 and SL_2 into Rx beamformer using the mechanism described in section 2.5. The delay information for SL_1 in frame 1 will also be loaded into the T_x beamformer inside the 128-channel Tx AFE (in Figure 16, $SL_{a,b}$ indicates the bth scanline in the ath frame). This configuration time is pre-determined, and this allows the Tx beamformer to send a "Ready" signal to the RTC. Upon receiving the signal, the RTC initiates the Tx pulse firing by sending pulse-in, pulse-enable, and scan-enable signals. After Tx firing, the Rx beamformer begins to listen and beamform received signals. The data is sampled based on the ADC sample rate. DDR3 control data has already been preloaded into FPGA local memory for SL_1, so this is applied to the coarse delay for beamforming. The I/Q data is then backhauled throught the PCIe solution based wireless module. Notice there is a gap named "Offset" in Figure 16. This offset time is used when the R_x beamformer sends beamformed data to the display device and the T_x beamformer receives the 3 signals (pulse-in, pulse-enable, and scan-enable) from the RTC. This offset time is the reason why the WBL is the lower bound for PRI, as mentioned in section 4.

The program on Android device, once the system starting up, will wirelessly send the pre-calculated delay information to the probe, which will send the Android device the digitized data pack during the wireless backhaul after applying the information and beamforming. Upon receiving the package, the program will create an image and apply a filter on it. The filtered image will finally be displayed.

The dynamic power management approach allows power reduction on a high end system with higher channel count and operating at higher supply voltages. The approach allows dynamic control of the AFE operational state based on the ultrasonic pulse repetition frequency (PRF). The PRF of an ultrasonic smart probe is dictated primarily by the speed of its wireless backhaul. Regardless of the depth of penetration, there is going to be a sizable offset after the successful reception of the furthest received signal, and the beginning of the next transmit firing. This offset region, as has been discussed above, is where the receiver is sitting in an idle state and burning DC power. Hence, we would prefer to turn off the most power inefficient regions of the receiver.

Our ultrasound imaging system supports 40 MHz ADC sample rate (25 ns temporal resolution) with 12 MHz max fundamental frequency and 80% bandwidth. Also, a max pass band of 16.8 MHz is available with 20 dB rejection. For the fine delay filter, each filter is implemented based on 48 FIR filter, which has 0.2 dB ripple and 20 dB rejection. The PRI we used is 200 μs with FR of 39 fps for 128-channels, limited by RTD of 65 μs at penetration depth of 5 cm, and a WBL of 180 μs (N_b = 12 bits/sample, W_b = 140 MHz). The TDL of 51.2 μs, constrained by the bus speed of 40 MHz. The digital beamformer is implemented with register-transfer level (RTL) in Verilog and synthesized by Quartus II (Altera Inc., San Jose, CA, USA) for Altera Arria V. The flow chart of the RTC is shown in Figure 17. The Android GUI program has been developed with C++ to control the system. All the control signals were generated in the tablet and are transferred wirelessly to the prototype board after the user configures the required input data via the tablet.

To improve the utilization of memory, a delta-coding scheme is used. From the previous mechanism, we know that the coarse delay for each channel monotonously increases and the maximum difference of coarse delay signals between two focal points for each channel is one. One bit is enough to be used as the delta between two coarse control signals. The absolute values of coarse delays for the first focal point are given at the beginning of scanning a line. The delays change according to the delta value. The fine delay has 3 bits, which are updated every sample. Therefore, there are 4 bits to be stored for each sample and each channel. Figure 18 shows the data organization in local memory. The memory is updated every scanline. The beginning 9 bits are used for initial absolute coarse delay. Subsequent 3 bits are used to indicate the fractional delay for the first focal point. Following 4-bit words are used to update delay information, whose first bit is the increment of coarse delay and other 3

bits are the fractional delay.

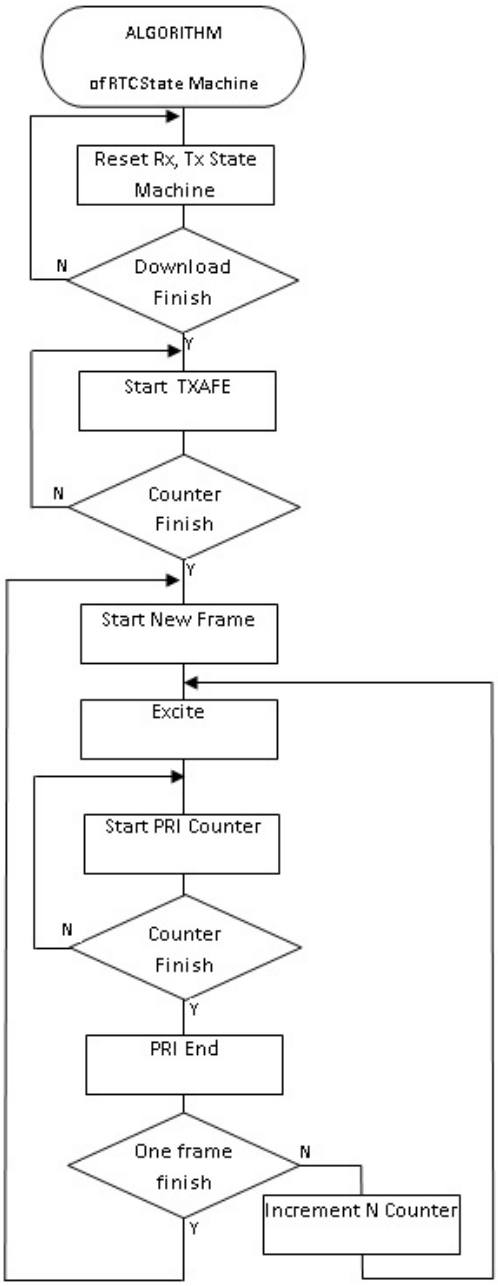


Figure 17: Flow chart of the real time controller for the ultrasound digital beamformer.

The system provides 128-channels of transmitter channels with a maximum of 32 excited pulses while simultaneously receiving signals at a 40 MHz sampling rate and converting them to a 12-bit binary number. The digitized data from all channels are first fed through the processor in the FPGA, and then stored in LPDDR3 memories by using direct memory access. These raw data are accessed by the beamforming processor to build the image that will be sent to the Android device wirelessly for further processing.

Figure 19 shows the finite state machine (FSM) used to coordinate the data flow of the local memory to dynamically generate the control signals (Coarse Delays and Fine Delays). Since there is limited bandwidth between the receiver and the memory, we introduced the scheme to dynamically change delays to accommodate the dynamic beamforming. To do that, last-level local memory is used on the receiver, which is 16×4×9 large for 16 channels. The memory is organized so that each channel has 4 × 9 bits in the local memory. Initially, those memories are filled by 4 × 9 bits of data, so that the initial fine delays and fraction delays are known (This is done in the first 4 status S0, S1, S4, S5). Once those 4 statuses are set, the delays can be initialized. Then the delays are held until the receiver starts working (state S2). As the receiver starts receiving data, the delays change. The new delay information (4 bits, 1 for coarse delay change, 3 for new fine delay) is sent into the control blocks in each cycle. In this phase, the local memory needs to keep updating data (S3 does this). When the counter hits the length of the scanline depth, one scanline is done. It then reverts to the initial state.

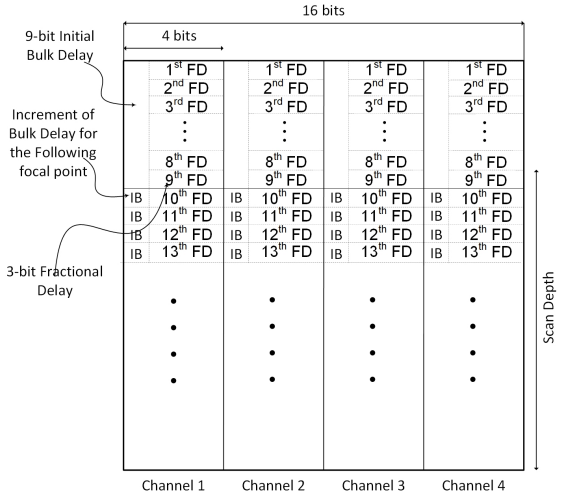


Figure 18: Flow chart of the real time controller for the ultrasound digital beamformer.

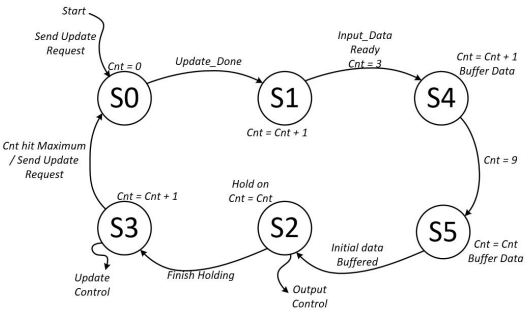
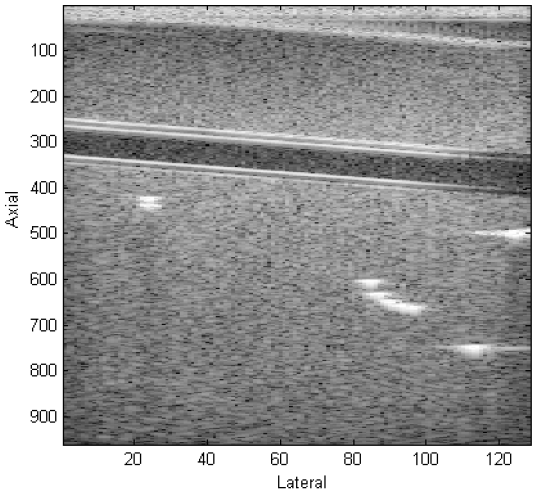


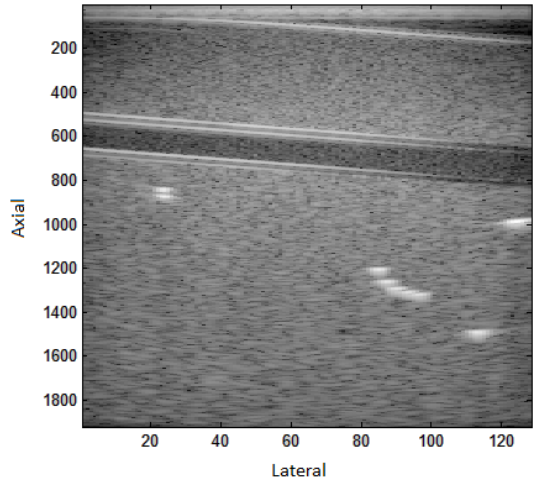
Figure 19: Finite state machine for generating the control signal for the ultrasound digital beamformer based on a single FPGA.

The beamformer that applies delays to the echoes of each channel is implemented by the strategy that combines coarse and fine delays (4 ns). This system is capable of achieving a maximum frame rate of 50 fps; this is mainly limited by the maximum speed we can achieve while wirelessly backhauling the data to the Android device. Table 3 shows the device utilization for the

whole architecture, mid-end and the control block. The 128-channel model consumes 130,547 adaptive logic modules (ALMs), the beamforming block consumes 128,044 ALMs, and the control section consumes 2,147 ALMs. The designed board contains the architecture in FPGA and DDR3, which is the external memory to store all delay information. We use the FPGA Mezzanine Card (FMC) connector to feed testing data into FPGA from stimulus boards. The board wirelessly backhauls the ultrasound image for display on a tablet or smart phone as shown in Figure 20(a).



(a)



(b)

Figure 20: Ultrasound images displayed on tablet processed by (a) single Altera Arria V FPGA based ultrasound digital imaging system and (b) Simulink.

To validate our design, another FPGA is added for testing purposes: the beamformed data is stored temporarily in the added FPGA and will be sent to the Android device after testing. A software-based beamformer is simulated by Matlab and Simulink. Data generated by a quantized Matlab module is used to verify the output of the board, as shown in Figure 20(b). The sum operation across all channels is pipelined to incorporate numerous inputs and process them at a high clock frequency. The first stage in the pipeline contains 64-inputs, each producing the sum of the incoming 12-bits from 2 channels. The resulting 32 outputs are summed in pairs in the second pipeline stage. The last summation stage is truncated to a 12-bit number and fed into the high pass filter. The image is generated from a Verasonics machine (Vantage 128, Verasonics Inc., Kirkland, WA, USA).

Table 3: Implementation Parameters

Parameter	Value	Percentage
ALMs	130,547	96
Registers	41,450	-
Pins	263	40
Block memory bits	2,754,067	18
PLLs	3	8
DLLs	1	25

7 Conclusion

In this paper, we present a series of techniques: interpolation filter, limited aperture size and dynamic power management to improve the efficiency of a fully dynamic ultrasound beamformer for a handhold real-time 128-channel ultrasound system, providing higher image resolution compared to the 32-channel [26] and 64-channel beamformer [32]. The proposed architecture is programmed and synthesized in a single Arria V FPGA using Verilog, allowing fully dynamic beamforming in a low-cost single FPGA implementation. This proposed digital beamformer can be used in generic ultrasound imaging systems, personal portable ultrasound health care, as well as the cart ultrasound machine. It is useful for portable low-power ultrasound and high performance ultrasound imaging with the genuine support of dynamic beamforming, while most current ultrasound devices employ pseudo dynamic beamforming. In addition, the power consumption and hardware utilization was reduced by a huge amount.

References

[1] J. A. Jensen, "Medical ultrasound imaging," *Progress in Biophysics and Molecular Biology*, vol. 93, pp. 153-165, no.1, 2007.

[2] J. J. Hwang, J. Quistgaard, J. Souquet, and L. A. Crum, "Portable ultrasound device for battle field trauma," in *Proc. IEEE Ultrason. Symp.*, vol. 2, pp. 1663-1667, 1998.

[3] M. Karaman, P. c. li, and M. odonnell, "Synthetic aperture imaging for small scale systems," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 42, no. 3, pp. 429-442, 1995.

[4] C. Yoon, J. Lee, Y.M. Yoo, T.-K. Song, "Display pixel based focusing using multi-order sampling for medical ultrasound imaging,", 3rd ed. *Electronics Letters*, vol. 45, no.25, pp. 1292-1294, Dec. 3, 2009.

[5] A. Assef, J. Maia, E. Costa "A flexible multichannel FPGA and PC-based ultrasound system for medical imaging research: initial phantom experiments," *Research on Biomedical Engineering*, vol. 31, pp. 277-281, Sept. 2015.

[6] A. Murtaza, D. Magee, and U. Dasgupta. "Signal processing overview of ultrasound systems for medical imaging," *SPRAB12, Texas Instruments*, Nov. 2008.

[7] J. A. Jensen, O. Holm, L. J. Jensen, H. Bendsen, H. M. Pedersen, K. Salomonsen, J. Hansen, and S. Nikolov, "Experimental ultrasound system for real time synthetic imaging," *IEEE Ultrasonics Symp*, pp. 1595-1599, 1999.

[8] M. O'Donnell, M. J. Eberle, D. N. Stephens, J. L. Litzza, B. M. Shapo, J. R. Crowe, C. D. Choi, J. J. Chen, D. W. M. Muller, J. A. Kovach, R. J. Lederman, R. C. Ziegenbein, K. San Vicente, and D. Bleam, "Catheter arrays: can intravascular ultrasound make a difference in managing coronary artery disease," *IEEE Ultrasonics Symp*, pp. 1251-1254, 1997.

[9] J. Schulze-Clewing, M. J. Eberle and D. N. Stephens, "Miniaturized circular array," *IEEE Ultrasonics Symp*, pp. 1253-1254, 2000.

[10] J. P. Stitt, R. L. Tutwiler and K. K. Shung, "An improved focal zone high-frequency ultrasound analog beamformer," *IEEE Ultrasonics Symp*, pp. 617-620, 2002.

[11] C. Dusa, P. Rajalakshmi, S. Puli, U. B. Desai and S. N. Merchant, "Low complex, programmable FPGA based 8-channel ultrasound transmitter for medical imaging researches," in *2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Natal, 2014, pp. 252-256.

[12] A. Bhuyan *etal*., "Integrated circuits for volumetrix ultrasound imaging with 2-D CMUT arrays," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, pp. 796-804, Dec. 2013.

[13] J. W. Choe, O. Oralkan, and P. T. Khuri-yakub, "Design optimization for a 2-D sparse transducer array for 3-D ultrasound imaging," *Proc. IEEE Ultrason*. Sysmp., 2010, pp. 1928-1931.

[14] J. Song *etal*., "Reconfigurable 2D cMUT-ASIC arrays for 3D ultrasound image," *Proc. SPIE Meidcal Imaging*, vol. 8320, p. 83201A, Feb. 2012.

[15] V. S. Gierenz, R. Schwann,T .G. Noll, "A low power digital beamformer for handheld ultrasound systems," in *Solid-State Circuits Conference, Proceedings of the 27th European*, pp. 261-264, 18-20 Sept, 2001.

[16] M. Almekkawy , J. Xu and M. Chirala, "An optimized ultrasound digital beamformer with dynamic focusing implemented on FPGA," in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 26, pp. 3296-3299, 2014.

[17] C. Fritsch, M. Parrilla, T. Sanchez, O. Martinez, "Beamforming with a reduced sampling rate," *IEEE Ultrasonics*, vol. 40, pp. 599-604, 2002.

[18] S. R. Freeman, M. K. Quick, M. A. Morin, R. C. Anderson, C. S. Desilets, T. E. Linnenbrink, and M. ODonnell, "Delta sigma oversampled ultrasound beamformer with dynamic delays," *IEEE Trans. Ultrason.*, Ferroelect., Freq. Contr., vol. 46, pp. 320-332, 1999.

[19] M. Kozak and M. Karaman, "Digital phased array beamforming using single-bit delta-sigma conversion with non-uniform oversampling," *IEEE Trans. Ultrason.*, Ferroelect., Freq. Contr., vol. 48, pp. 922-931, 2001.

[20] C. Basoglu, R. Managuli, G. York, and Y. Kim, "Computing requirements of modern medical diagnostic ultrasound machines," *Parallel Computing*, vol. 24, pp. 1407-1431, 1998.

[21] C. H. Hu, X. C. Xu, J. M. Cannata, J. T. Yen and K. K. Shung, "Development of a real time high frequency ultrasound digital beamformer for high frequency linear array transducers," *IEEE Trans. Ultrason.*, Ferroelect., Freq. Contr., vol. 53, no. 2, pp. 317-323, 2006.

[22] D. K. Peterson and G. S. Kino, "Real-time digital image reconstruction: A description of imaging hardware and an analysis of quantaization errors," *IEEE Trans. Sonics Ultrason*, vol. SU-31, pp. 337-351, 1984.

[23] B. D. Steinberg, "Digital Beamforming in ultrasound," *IEEE Trans. Ultrason.*, Ferroelect., Freq. Contr., vol. 39, no. 2, pp. 716-721, 1992.

[24] H. Sohn, S. Seo, J. Kim and T. Song, "Software implementation of ultrasound beamforming using ADSP-TS201 DSPs," *Proceedings of SPIE*, vol. 6920, pp. 69200Z-69200Z-11, 2008.

[25] H. T. Feldkamper, R. Schwann, V. Gierenz, T.G. Noll, "Low power delay calculation for digital beamforming in handheld ultrasound systems," *IEEE Ultrasonics Symposium*, vol. 2, no., pp. 1763-1766, Oct 2000.

[26] Gi-duck Kim, Changhan Yoon, Sang-Bum Kye, Youngbae Lee, Jeeun Kang, Yangmo Yoo, Tai-Kyong Song, "A single FPGA-based portable ultrasound imaging system for point-of-care applications," *IEEE Trans. Ultrason.*, Ferroelect., Freq. Contr., vol. 59, no. 7, pp. 1386-1394, July 2012.

[27] B. Tomov and J. A. Jensen, "A new architecture for a single chip multi-channel beamformer based on a standard FPGA," *IEEE Ultrasonic Symp.*, pp. 1529-1533, 2001.

[28] P. J. Cao and K. K. Shung, "Design of a real time digital beamformer for a 50 MHz annular array transducer," *IEEE Ultrasonic Symp.*, pp. 1619-1622, 2002.

[29] P. J. Cao,C. H. Hu and K. K. Shung, "Devolepment of a real time digital high frequency annular array ultrasound imaging system," *IEEE Ultrasonics Symp*, pp. 1867-1870, 2003.

[30] A. A. Assef, J.M. Maia, F.K. Scheneider, E.T. Costa, V.L. Button, "A programmable FPGA-based 8-channel arbitrary waveform generator for medical ultrasound research activities," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, vol., no., pp. 515-518, Aug. 28 2012-Sept. 1, 2012.

[31] M. A. Hassan, and Y. M. Kadah, "Digital signal processing methodologies for conventional digital medical ultrasound imaging system," *American Journal of Biomedical Engineering*, vol. 3,no. 1, pp. 14-30, 2013.

[32] C. Hu, L. Zhang, J. M. Cannata, J. Yen, K. K. Shung, "Development of a 64 channel ultrasonic high frequency linear array imaging system," *Ultrasonics*, vol. 51, no. 8, pp. 953-959, 2013.

[33] A. A. Lazar and L. T. Toth, "Perfect recovery and sensitivity analysis of time encoded bandlimited signals," *IEEE Trans. Circuits Systems 1*, vol. 51, no. 10, pp. 2060-2073, Oct. 2004.

[34] N. S. Jayant and A. E. Rosenberg "The preference of slope overload to granularity in the delta modulation of speech," *Bell System Technical Journal*, vol. 50, no. 10, Dec. 1971.

[35] A. S. Alvarado and J. C. Principle, "From compressive to adaptive sampling of neural and ECG recordings," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 633-636, Apr. 2011.

[36] J. Lu, H. Zou, and J. F. Greenleaf, "Biomedical ultrasound beam forming," *Ultrasound in Med. and Biol.*, vol. 20, pp. 403-428, 1994.

[37] H. Yao, "Synthetic aperture methods for medical ultrasonic imaging,", Ph.D. thesis, Univ. of Oslo, Norwa, 1997.

[38] I. LIE, M. E. TANASE, D. LASCU, and M. LASCU, "Ultrasonic beamforming with delta-sigma modulators," in *Proc. of the 10th WSEAS International Conference on CIRCUITS*, 2006, pp. 344-349.

[39] R. A. Mucci, "A comparison of efficient beamforming algorithms," *IEEE Trans. Acoust. Speach Signal Processing*, vol. ASSP-32, no. 3, pp. 548-558, 1984.

[40] T. I. Laasko, V. Valimaki, M. Karjalainen, and U .K. Laine "Splitting the unit delay," *IEEE Signal Processing Mag.*, vol. 13, pp. 30-60, 1996.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 61-66 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Effects of Dielectric Properties of the Material located inside Multimode Applicator on Microwave Efficiency

Sofiya Ali Mekonnen, Sibel Yenikaya*, Gökhan Yenikaya, Güneş Yılmaz

Department of Electrical-Electronics Engineering, Uludag University, 16059, Turkey.

ARTICLE INFO

Article history:
Received: 27 March, 2018
Accepted: 27 April, 2018
Online: 20 May, 2018

Keywords:
Microwave heating
Efficiency
Electric field distribution
Dielectric property
Clay sample

ABSTRACT

During microwave heating of materials, the efficiency of microwave heating depends on the materials' dielectric property, shapes and sizes of the material, materials' position inside the applicator, operating frequency, level of input power, specific heat capacity, number and position of waveguide over the applicator, size and geometry of applicator etc. This paper examines the effects of dielectric properties of the clay sample placed in the multimode applicator on the performance of microwave heating. The simulated and experimentally obtained results show that the variation in clay samples dielectric value creates variation in system efficiency inside the microwave applicator. Also placing another dielectric slab over the material to be heated affects the electric field distribution and system efficiency. Placing KESTAMID dielectric slab over the material improved the heating efficiency by 22%. COMSOL Multiphysics software was used to simulate and estimate the electric field distribution over the surface of the clay sample and inside the multimode applicator for different dielectric property clay samples. The simulated and experimentally obtained results are almost matched.

1. Introduction

In recent days, microwave heating is gaining popularity as an alternative heating technique in many industries due to its high heating rates, significant energy saving, reduced processing time, selective and volumetric heating. Also microwave applications involve environmental friendly usage, safer handling and improved quality of materials etc [1,2]. Due to its various advantages, microwave has found applications in the drying of foods [3], drying of textiles[4], drying of zinc sulfate [5], recycling of polymeric materials [6], coal mining [7], vulcanizations of rubber[8,9], drying and sintering of ceramics [10], and drying of bagasse [11]. However, low heating efficiency is a limitation for microwave heating applications in practice.

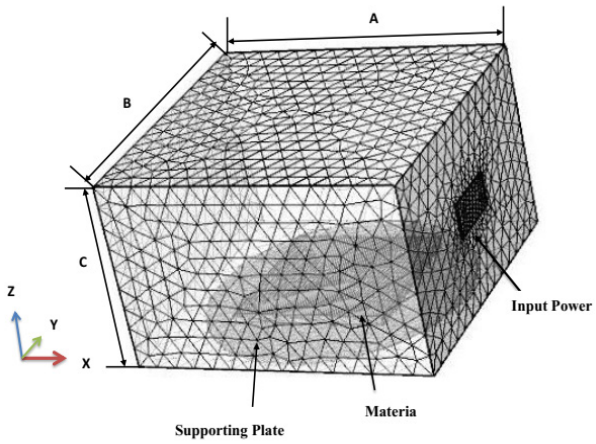
High heating efficiency is one of the most desired characteristics for both industrial and domestic microwave ovens since electric energy is expensive and because low efficiencies indicate high power reflections, which might damage the magnetron. Also the quality of the processed material is determined by its heating efficiency. So far microwave heating efficiency is improved by matching the load through different methods like using external devices, such as tuners, stubs or irises, feeding the applicator from more than one position [12], covering the sample material with dielectric multilayer [13] or by changing the position of the sample material inside multimode applicator [14]. However, the above mentioned techniques have their own limitation and difficulties in the overall cost of the system.

In microwave heating process, a number of factors are responsible for electric field distribution inside multimode microwave applicator. Among them samples' location, dielectric property, sample size and shape, applicators' size and its geometry, incident power and operating frequency are most important. In recent years, the effects of size, shape, applied power, operating frequency, samples' location and dielectric properties have been studied both experimentally and numerically on microwave heating efficiency and it was found out that for each different sample, varied efficiency is obtained [15-18].

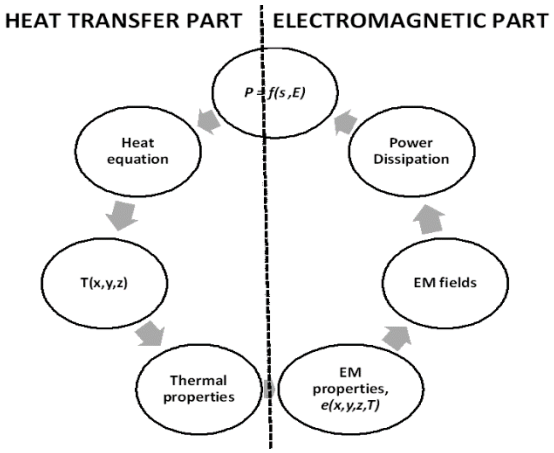
This paper is an extension of work presented in ELECO 2017 conference concerning the effects of dielectric properties of the clay sample placed in the multimode applicator on the performance of microwave heating [19]. The main goal of the extended version is to investigate the effects of dielectric property on the performance of microwave heating by increasing the number of sample dielectric materials and by placing a dielectric slab over the

*Corresponding Author: Sibel Yenikaya, Email: sguler@uludag.edu.tr

clay sample to investigate the effect of dielectric property on the microwave heating performance in the multimode applicator. The electric field distribution is carried out by COMSOL Multiphysics commercial software to investigate the effects of the dielectric properties.



a) Microwave oven modelling



b) Microwave- heat transfer model

Figure 1. Microwave oven modelling and microwave-heat transfer model

2. Theoretical study

The schematic and modeling diagram of microwave oven prepared for this study is shown in Figure 1. The microwave multimode applicator is represented by a metallic box of rectangular shape, (A = 354 cm, B = 337mm and C = 215 cm), typical of many industrial and microwave multimode applicator. The excitation port is a rectangular waveguide and the fundamental mode is TE_{10}, which operates at 2.45 GHz frequency. The sample is placed over a cylindrical shaped glass plate at z = 10mm, which is located near to the bottom of the applicator. For this particular study rectangular shape clay sample was chosen. The results obtained in this work can be applied in industrial processes where pieces of clay must be dried. Figure 1a shows schematic diagram of microwave oven, the computational domain was meshed using a tetrahedral element. The complete mesh consists of 239920 domain elements, 25550 boundary elements and 831 edge elements. Figure 1b shows microwave and heat transfer model, which is the relation between electromagnetic field and temperature distribution. The heating source is provided by the microwave dissipated power. Also the temperature variation during microwave heating process causes variations in electromagnetic properties. In this study, the distribution of electromagnetic equation inside the cavity and sample has been solved by the well-known Maxwell's equations, which can be reduced, in the frequency domain with the aid of the following vector-wave equation [1];

$$\nabla^2\vec{E} + \omega^2\mu\varepsilon\vec{E} = 0 \quad (1)$$

where $\vec{E}$ is the vector's electric field, ω is the angular frequency, μ is the magnetic permeability, and ε is the dielectric complex permittivity of the medium, given by [2];

$$\varepsilon = \varepsilon_0(\varepsilon' - j\varepsilon'') = \varepsilon_0(\varepsilon' - j\frac{\sigma}{\omega}) \quad (2)$$

where ε_0 is the vacuum permittivity, ε' is the dielectric constant, ε'' is the loss factor, and σ is the conductivity of the medium. The ability of a dielectric material to absorb microwaves and store energy is given by the complex permittivity of the medium.

The ratio of the dielectric loss to the dielectric constant is known as the loss tangent (tan δ) which is given as;

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} \quad (3)$$

Hence with values of less dielectric constant and large values of loss tangent or dielectric loss, materials couple with microwave with great efficiency. In addition, the dielectric properties of a material depend upon the temperature, frequency and the composition of the material.

The efficiency of the system is calculated as, the ratio of the absorbed power by the material to be heated (P_b) to the microwave input power (P_a);

$$\eta = \frac{P_a}{P_b} \times 100 \quad (4)$$

The above mentioned equations with the proper boundary conditions of the metallic walls and the waveguide excitation are used to find the electric field distribution and heating efficiency during microwave heating.

3. Results and discussion

The above mentioned equations have been applied to study the effects of different dielectric properties of clay samples placed inside multimode microwave applicator as shown in Figure 1b for efficient microwave heating. The sample is heated for 5 minutes at 2.45 GHz operating frequency inside the multimode microwave applicator with initial temperature of 9 ℃. The applicator is fed 900 W for each simulation works through the rectangular waveguide. The sample is placed over a cylindrical shaped glass plate, which is located near to the bottom of the applicator. The dielectric values used for each simulation and experimental work is measured by network analyzer (Agilent, E506 1B) instrument and two parallel 200mm probe at 2.45 GHz frequency. Table 1 show the samples' dielectric permittivity used in all the simulation and experimental works.

Table 1. Dielectric characteristics of the clay samples

	ε'	tan δ (10^{-3})
Sample 1	90	47.22
Sample 2	77.10	119.19
Sample 3	59.99	187.19
Sample 4	52.60	239.35

For this particular study, first, we used four different dielectric property values as shown in table 1 to investigate the effects of dielectric properties on the performance of microwave heating. Later, we investigated the influence of additional dielectric property value located inside multimode applicator on performance of microwave heating, which is located at the top side of sample material by using KESTAMID dielectric slab of 1 cm thickness. KESTAMIDs' relative permittivity is εr = 3.7.

Microwave heating efficiency is greatly affected by geometry and size of the sample, its dielectric properties, operating frequency, applied power, its placement inside the applicator, size and geometry of the applicator. However, in this study we will observe and discuss only the effects of the samples' dielectric properties while keeping the other factors constant for the entire simulation and experimental work.

The dielectric properties of the samples' located inside the applicator influences the amount of the microwave power absorbed by the sample during its heating process. Dielectric properties of the material to be heated are functions of temperature, frequency and the composition of the material. The change of the dielectric properties with temperature influences the standing wave pattern inside the cavity and thus the microwave absorption. However, for this study we kept the effects of temperature and frequencies on dielectric properties of the sample constant for the purposes of simplicity in electromagnetic field and related calculations.

Figure 2a shows the electric field distribution over the clay sample 1 during microwave heating process for rectangular shaped sample. The simulation result show that the electric field intensity is stronger near to the top and bottom walls of the microwave applicator, and the weakest is located in the middle of the microwave applicator. This variation in electric field distribution leads to different microwave efficiency. From the simulation, the obtained absorbed power by the sample during microwave processing is 335.93 W. From this we calculated the efficiency of the system as η =37.33%.

To validate the simulated electric field distribution of clay samples, experimental works had been done to measure temperature distribution. A thermal fax paper was placed on the top of the sample before microwave heating and placed inside the applicator. Figure 2b shows the stain made by clay sample 1 during microwave heating. The stains obtained on the fax paper agree with the hot and cold spots generated during microwave heating. The temperature pattern over the samples was measured by the laser thermometer as shown in figure 2c.

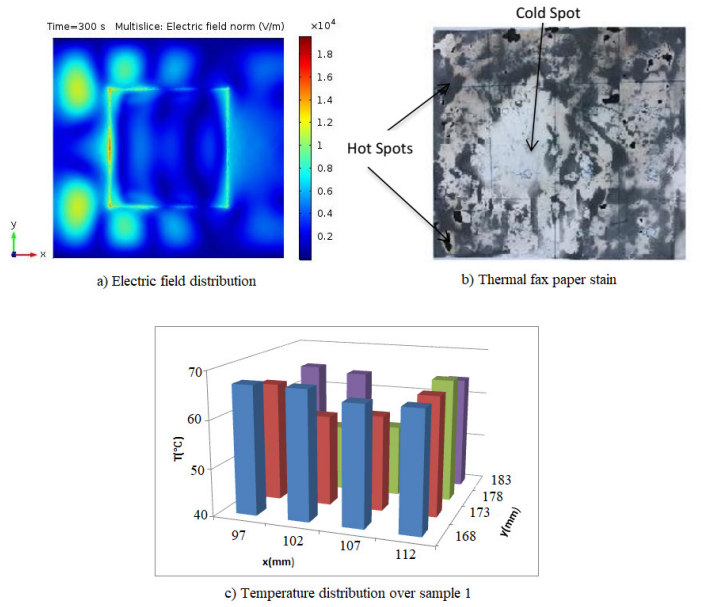


Figure 2. Electric field distribution, Thermal fax paper stain and temperature distribution of sample 1 (ε_r = 90-j4.25) during microwave drying.

Figure 2c represents temperature distribution pattern for sample 1, generally coincides with the simulated electric field distribution. For example, at the region of maximum electric field intensity on sample 1, V= 23.193 kV/m and 28.521 kV/m, the temperature that rises over sample 1 is also at the maximum; i.e., 61℃ and 67℃, respectively. Also for the other samples it shows that, at maximum electric field intensity region, the temperature rise is also maximum and at lower electric field region, the temperature rise is low. In the middle zone of the multimode applicator, the temperature rises of sample materials are lower. From the above experimental result, we can see clearly the direct relationship between the electric field intensity and temperature pattern, the temperature rise is affected by the sample's dielectric properties, which also affected the system's overall efficiency, and this validates our work.

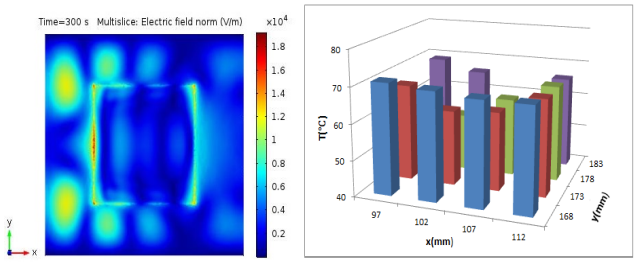
Figure 3 shows the electric field and temperature distribution over the clay samples and inside multimode applicator during microwave heating process for rectangular shaped samples. Figure 3a shows electric field distribution over sample 2. The power absorbed by the sample is 380.34 W. From here the efficiency of the system is η = 42.31%. Figure 3b shows the temperature distribution pattern over sample 2. After the microwave heating, the minimum and maximum temperature over the heated material was measured as minimum temperature Tmin = 53 ℃ and maximum temperature Tmax=71 ℃. The average temperature distribution was calculated as 62 ℃.

Figure 4 shows the electric field and temperature distribution over the clay samples and inside multimode applicator during microwave heating process for rectangular shaped samples. Figure 4a shows electric field distribution over sample 3. The power absorbed by the sample 3 is 430.83 W. From this the calculated efficiency of the sample is 47.87%. Figure 4b shows the temperature distribution pattern over sample 3. After the microwave heating, the minimum and maximum temperature over the heated material was measured as minimum temperature Tmin

= 55℃ and maximum temperature Tmax=71℃. The average temperature distribution was calculated as 63 ℃.

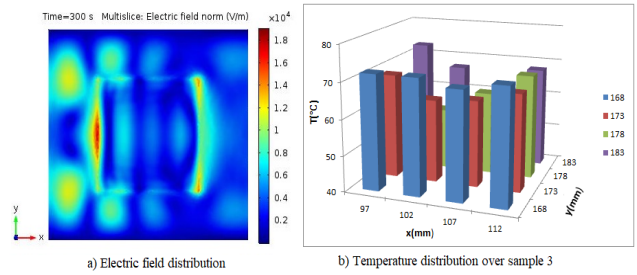
Figure 5 shows the electric field and temperature distribution over the clay samples and inside multimode applicator during microwave heating process for rectangular shaped samples. Figure 5a shows electric field distribution over sample 4. Figure 5a shows the electric field distribution over sample 4. The power absorbed by sample 4 is 454.91 W, and from here the calculated system efficiency is 50.55%. Figure 5b shows the temperature distribution pattern over sample 4. After the microwave heating, the minimum and maximum temperature over the heated material was measured as minimum temperature Tmin = 65℃ and maximum temperature Tmax=75℃. The average temperature distribution was calculated as 70 ℃.

Figure 6 shows summary of the results, obtained absorbed power and system efficiency for different dielectric values of rectangular clay samples mentioned in table 1.



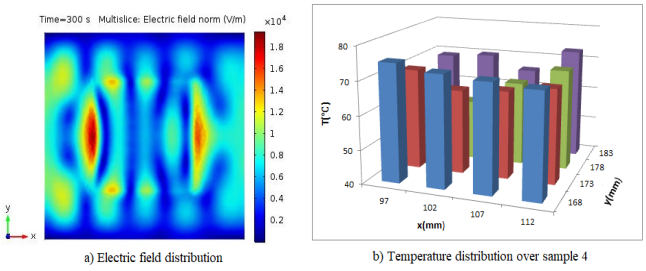
a) Electric field distribution b) Temperature distribution over sample 2

Figure 3: Electric field and temperature distribution over sample 2



a) Electric field distribution b) Temperature distribution over sample 3

Figure 4: Electric field and temperature distribution over Sample 3



a) Electric field distribution b) Temperature distribution over sample 4

Figure 5: Electric field and temperature distribution over Sample 4

From the above simulated and calculated results, under assumption of similar initial temperature, size and geometry of the sample, samples' placement inside applicator, frequency, and applicators' size and geometry, the electric field distribution and efficiency of the system is different for different samples' dielectric property.

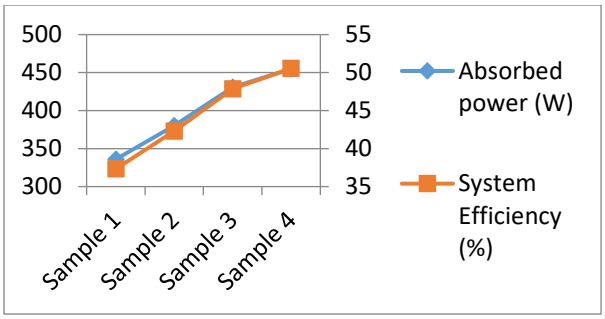


Figure 6. Absorbed power and system efficiency for different clay sample.

Secondly in this extended version, we placed a dielectric slab over the clay sample to see its dielectric property influence on the performance of microwave heating as shown in figure 7. Figure 7 shows the schematic diagram of microwave oven loaded with KESTAMID slab and clay sample.

When an additional dielectric slab is placed around the sample material, the original electric field distribution inside multimode applicator changes, because between the two different materials reflection and transmission of electromagnetic field occurs thus it leads to variation in systems efficiency. In this study, we used rectangular shaped KESTAMID dielectric slabs of 1 cm thickness to place over the clay sample to investigate the influence of adding another dielectric property on the performance of microwave heating inside multimode applicator.

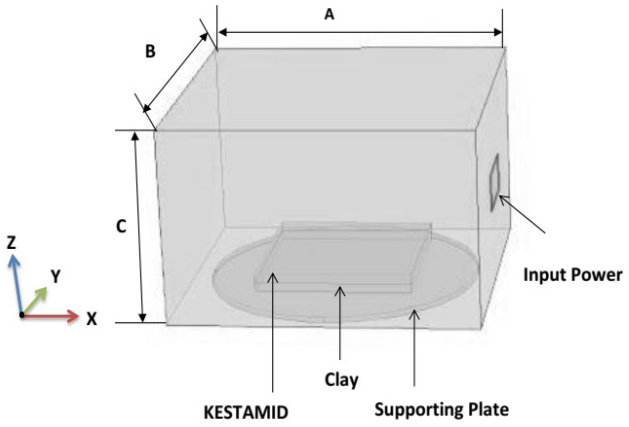
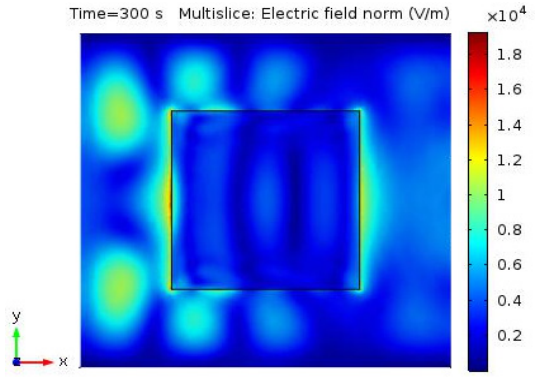


Figure 7: Schematic diagram of loaded microwave oven with KESTAMID slab and clay sample.

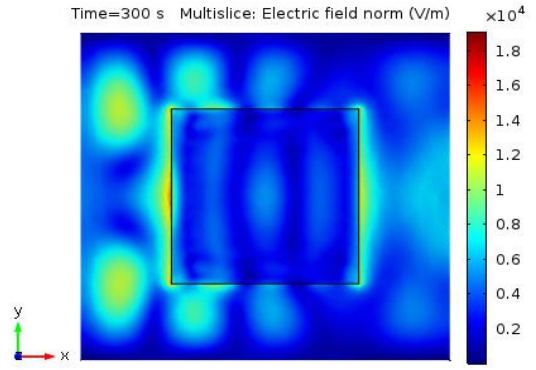
Figure 8 shows the electric field distribution over clay sample and inside multimode applicator after KESTAMID slab is placed over clay samples having different dielectric property.

Figure 8a shows the electric field distribution over sample 1. From the results obtained from simulation the absorbed microwave power by sample 1 during microwave processing is 464.21 W. From this we calculated the efficiency of the system as η = 51.58%. Figure 8b shows electric field distribution for clay sample 2. The power absorbed by sample 2 is 504.61 W. From this the efficiency of the system is η=56.07%. Figure 8c shows electric field distribution for clay sample 3. The power absorbed by sample 3 is 561.26 W. From this the calculated efficiency of the sample is 62.36%. Figure 8d shows the electric field distribution over sample 4. The power absorbed by sample is 649.09 W, and from here the efficiency of system is 62.19%. Figure 9 shows summary of the

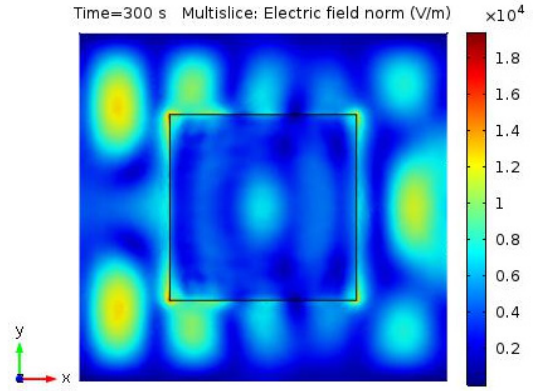
a) Sample 1



b) Sample 2



c) Sample 3



d) Sample 4

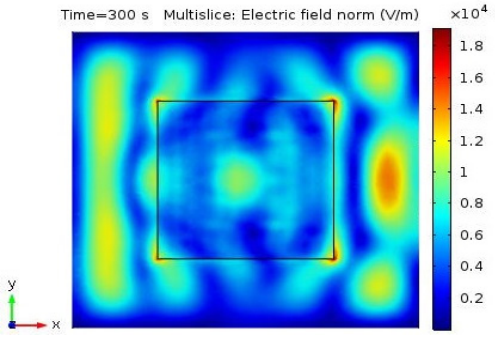


Figure 8: Electric field distribution inside multimode applicator during dielectric slab placement over the clay sample; a) sample 1 b) sample 2 c) sample 3 d) sample 4.

obtained results for different dielectric values of rectangular clay samples when KESTAMID slab is placed over the material to be heated.

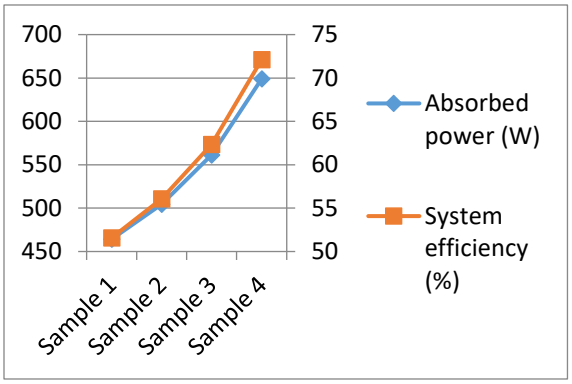


Figure 9. Absorbed power and systems' efficiency of different clay samples (KESTAMID slab on top of clay sample).

As shown in Figure 8 and relating numerical analysis, we can see that the electric field distribution and efficiency of the system is greatly affected by the newly added dielectric slabs' dielectric property. The electric field distribution shows more uniformity in Figure 8 compared to Figure 2,3,4 and 5. However, in Figure 8 the electric field distribution over clay sample materials having high loss tangent values shows more uniformity compared to materials having low loss tangent values. Also placing KESTAMID slab improves system efficiency by 22% compared to the efficiencies obtained without placing a dielectric slab.

Finally, from the analysis of the influence of the clay sample employed in the microwave heating process, it can be concluded that different samples' dielectric permittivity produce different levels of efficiency. Also by placing a dielectric slab over the material to be heated can improve systems efficiency. Therefore, the dielectric properties of the material located inside multimode applicator is of utmost importance when trying to design an efficient microwave oven.

4. Conclusion

In this paper, the effects of dielectric properties of clay samples on the performance of microwave heating have been investigated and discussed. Microwave heating is affected by different parameters like the placement of a material inside multimode applicator, size and shape of the material, operating frequency, dielectric properties of the material, size and shape of the applicator, and numbers and positions of the waveguide over the applicator. However, in this paper only the dielectric properties of the material have been studied. This has been done by keeping constant the other factors throughout the work in order to see solely the effect of the dielectric properties changes on system efficiency. Different clay samples' dielectric property gives different system efficiency. From the obtained simulation and experimental results, the materials with high loss tangent values give maximum system efficiency compared to the others. In other hand placing another dielectric slab (in this case KESTAMID) over the clay sample improves both electric field distribution pattern and system efficiency. Placement of KESTAMID slab over the material to be heated improved the efficiency of the system by 22%. Also a material with high loss tangent value shows more uniform electric

field distribution compared to the others. However, different dielectric slab thickness also has an effect on the performance of the microwave heating; it will be considered in the future works.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are thankful to Assoc. Prof. Dr. Güray SALİHOĞLU of Uludag University (Department of Environmental Engineering) and Assoc. Prof. Dr. Filiz ALTAY of Istanbul Technical University.

References

[1] A. C. Metaxas, R. J. Meredith, Industrial Microwave Heating, Peter Peregrinus, 1983.

[2] David K. Cheng, Field and wave electromagnetics. John Wiley & Sons, Inc., 2009.

[3] H. Feng, Y. Yin, J. Tang, "Microwave Drying of Food and Agricultural Materials: Basics and Heat and Mass Transfer Modeling". Springer-Verlag, Food Eng Rev, 4(2), pp. 89-106, 2012. https://doi.org/10.1007/s12393-012-9048-x

[4] J. Vrba, M. Stejskal, M. Pourova, "Microwave drying machine for textile materials", in proceedings of the 2009 European Microwave Conference (EuMC), Rome, Italy, pp. 1121-1124, 2009. IEEEXplore, DOI: 10.23919/EUMC.2009.5296116.

[5] L, A Jermolovicius, J. T. Senise, R. B. D. Nascimento, "Microwave Drying of Zinc Sulfate", in 2007 SBMO/IEEE MTT-S International Microwave & Optoelectronics Conference, Brazil, pp. 284-286, 2007. IEEEXplore, DOI: 10.1109/IMOC.2007.4404264

[6] G. A. Morozov, O. G. Morozov, R. R. Samigullin, A. R. Nasibullin, "Application of Microwave Technologies for Increase of Efficiency of Polymeric Materials Recycling", 2011 VIII International Conference on Antenna Theory and Techniques (ICATT), Kyiv, Ukraine, pp. 321-323, 2011. IEEEXplore, DOI: 10.1109/ICATT.2011.6170771

[7] Y. Hong, Bai. Lin, H. Li, H. Dai, C. Zhu, H. Yao. "Three-dimensional simulation of microwave heating coal sample with varying parameters", Applied Thermal Engineering, Vol. 93, pp. 1145–1154, 2016. https://doi.org/10.1016/j.applthermaleng.2015.10.041.

[8] N. Makul, P. Rattanadecho, "Microwave pre-curing of natural rubber-compounding using a rectangular wave guide", International Communications in Heat and Mass Transfer, Vol. 37, pp. 914–923, 2010. https://doi.org/10.1016/j.icheatmasstransfer.2010.03.001.

[9] U. Narumitbowonkul, P. Keangin, P. Rattanadecho, "Numerical and Experimental Analysis of Temperature Distribution and Electric Field in a Natural Rubber Glove during Microwave Heating", International Science Index, Mechanical and Mechatronics Engineering, 9(3), pp. 518-524, 2015. DOI: 10.1999/1307-6892/10000910.

[10] A. Esin, N. Mahmutyazıcıoğlu, S. Altıntaş, "Drying and Sintering of Ceramic Based Parts Using Microwave Heating", Key Engineering Materials, Vols. 264-268, pp. 731-734, 2004. https://doi.org/10.4028/www.scientific.net/KEM.264-268.731.

[11] S. K. Shah, M. A. Joshi, "Application of Microwave Heating In Vacuum for Bagasse Drying". Proceedings of the 2008 International Conference on Recent Advances in Microwave Theory and Applications, Jaipur, India, pp. 477-479, 2008. IEEEXplore, DOI: 10.1109/AMTA.2008.4763153.

[12] D. S. Rajpurohit, R. Chhibber, "Design Optimization of Two Input Multimode Applicator for Efficient Microwave Heating", International Journal of Advances in Microwave Technology (IJAMT), 1(3), pp. 68-73, 2016.

[13] J. Monzo-Cabrera, J. Escalante, A. Diaz-Morcillo, A. Martinez-Gonzalez, D. Sanchez-Hernandez, "Load matching in multimode microwave-heating applicators based on the use of dielectric layer moulding with commercial materials", Microwave Opt Technol Lett, 41(5), pp. 414–417, 2004. https://doi.org/10.1002/mop.20156.

[14] M. E. Requena-Perez, J. L. Pedreno-Molina, M. Pinzolas-Prado, J. Monzo-Cabrera, A. Diaz-Morcillo and D. Sinchez-Hernandez, "Load matching in multimode microwave-heating applicators by load location optimization", Proceedings of the 34th European Microwave Conference, , Amsterdam, The Netherlands, IEEE, Vol. 3: pp. 1549–1552, 2004. IEEEXplore, DOI: 10.1109/EUMC.2004.183902.

[15] J. L. Pedreno-Molina, J. Monzo-Cabrera, M. Pinzolas, "A new procedure for power efficiency optimization in microwave ovens based on thermographic measurements and load location search", International Communications in Heat and Mass Transfer, 34(5), 564-569, 2009. https://doi.org/10.1016/j.icheatmasstransfer.2007.02.002

[16] I. A. Ali, "Effect of load on the heating efficiency and temperature uniformity in multimode cavity applicators", Taylor & Francis Group, journal of microwave power and electromagnetic energy, 50(2), pp. 123-137, 2016. http://dx.doi.org/10.1080/08327823.2016.1190170.

[17] G. Brodie, "The influence of load geometry on temperature distribution during microwave heating", Transactions of the ASABE (American Society of Agricultural and Biological Engineers), 51(4), pp. 1401-1413, 2008. DOI: 10.13031/2013.25224.

[18] W. Cha-um, P. Rattanadecho, W. Pakdee, "Experimental analysis of microwave heating of dielectric materials using a rectangular wave guide (MODE: TE10) (Case study: Water layer and saturated porous medium)", Experimental Thermal and Fluid Science, 33(3), pp. 472–481, 2009. https://doi.org/10.1016/j.expthermflusci.2008.11.008

[19] S. A. Mekonnen, S. Yenikaya, G. Yenikaya, G. Yılmaz, "Effects of sample's dielectric property on the performance of microwave heating", proceedings of 2017 10th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey, pp. 1440 – 1443, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 64-71 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Novel MICS Receiver with FSK Dual Band Demodulator

Mouna Bettaieb*, Saif Benali, Ghazi Bouzid, Hatem Trabelsi

METS research Group, National Engineering School of Sfax, University of Sfax, BP 1173, 3038 Sfax, Tunisia.

ARTICLE INFO

Article history:
Received: 23 July, 2018
Accepted: 31 August, 2018
Online: 18 September, 2018

Keywords:
MICS receiver
Dual band FSK demodulator
Implantable medical device

ABSTRACT

A low-complexity dual-band chirp FSK, direct conversion receiver is described in this paper. The receiver is dedicated to be used in the transceiver unit of a medical implantable wireless sensor. The system uses the RF band between 402 and 405 MHz. Two sub-bands frequencies employing chirped pulses are assigned for both binary information. The novelty of this work is the use of a Binary FSK LFM modulator, a direct conversion receiver and a simple and low power non-coherent BFSK envelope detection demodulator. Receiver performances are evaluated for all the input power dynamic range. Receiver front-end parameters are optimized using harmonic balance simulation. In order to improve receiver sensitivity, a low pass filter with controllable bandwidth between 40 and 300 KHz is used to avoid in-band interference. The receiver is able to achieve a noise figure of 5.5 dB, a receiver sensitivity of -93 dBm and a maximum data rate of 100Kbps. The simulated IIP3 and P_{-1dB} are 12.6 dBm and 22.1 dBm respectively. A simple non coherent binary dual band FSK demodulator was used which is based on an envelope detector, integrate & dump, a sampling & hold and a liming circuit. The receiver was co-simulated with the dual band non coherent demodulator. The proposed receiver has a sensitivity of -93 dBm and a BER less than 10^{-3}.

1. Introduction

With the proliferation of portable or implantable devices in wireless healthcare sensor network, energy-efficient radios have become an active area of research.

This paper is an extension of work originally presented in International Conference on Engineering & MIS (ICEMIS 2017) [1].

Today, there is a great demand for low power transceivers dedicated to implanted sensors, thanks to the development of new technologies in microelectronics and the progress of sensor implantation techniques in the human body [2]. This system measures certain physiological parameters of people suffering from chronic diseases and communicates useful information to an information collection center in order to warn medical professionals in case of danger [3-5].

The implanted sensor node must provide continuous operation under battery power for at least 5 years.

WMTS (Wireless Medical Telemetry Services) and MICS (Medical Implantable Communications Services) are two standards used for medical electronic devices. Non implantable devices uses WMTS whereas MICS is dedicated to sensor networks implanted in the human body.

According to the MICS regulation from the Federal Communications Committee (FCC) its frequency band is between 402 MHz and 405 MHz which is divided into 10 channels. Each channel occupies a bandwidth of 300 KHz. The wireless communication range is limited to 2 m and the maximum transmitted power is set at -16 dBm [6]. These frequencies were selected for the following reasons: reasonable antenna size, good propagation characteristics for medical implants in the human body and worldwide availability [4].

In addition, MICS signals are safe for the human body thanks to their low power level. Moreover, this introduces low interference with other users. With this technology we can reach reduced complexity, low power and low cost transceiver.

Figure 1 introduces the block diagram of a typical sensor node [7-8]. Physiological signals are very weak and coupled with noise like body pressure or the amount of glucose in the blood of the patient. Thus, amplification and filtering process are required to increase the signal level and reject the noise. A multiplexer is used to switch between different sensors. The analog signal conversion into digital is ensured by an ADC. The microcontroller processes

*Mouna Bettaieb, National Engineering School of Sfax, BP 1173, 3038 Sfax, Tunisia, bettaiebmouna3@gmail.com

the signals from the sensors unit and executes the communication protocol. The MICS transceiver communicates the formatted data across a wireless link with a portable personal healthcare centre for testing and diagnosis by doctors.

A sensor node can perform communication in both directions. In transmission mode the sensor node sends all the useful sensors data and information regarding its operating status. In receive mode, it receives node configuration, wake up or standby information. It should also be noted that power consumption of a sensor node depends on the duty cycle, the time during which the system is active over a given period.

The use of CMOS technology in healthcare implanted devices is justified since the principal challenge is the minimization of the energy consumption in the sensor nodes in order to integrate the digital and the analog parts together on the same chip. That's why the architecture of the RF wireless sensors, modulation scheme and parameters optimization of different blocks of the transceiver was subjects of many studies and researches as part of the low power consumption challenge [9].

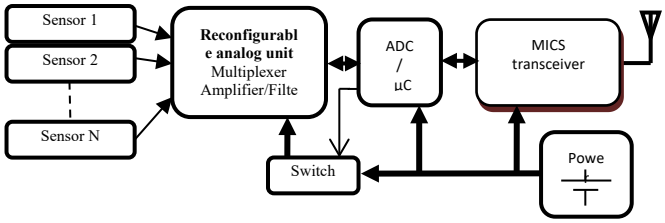


Figure 1. A typical sensor node.

Linear Frequency Modulation (LFM) is widely used in radars, but it has not been yet used for implanted sensor networks in the human body using the MICS band.

The purpose of this paper is to design a simple and low power, 402-405MHz-MICS receiver chain based on the LFM technique with a non-coherent BFSK envelope detection demodulator dedicated for implantable sensor node.

The description of Linear Frequency Modulation technique and transceiver architecture is presented in section 2. In section 3 the receiver design and simulations results will be presented. A co-simulation of the analog receiver with the non coherent dual band FSK demodulator will be shown in section 4. Section 5 presents blocks parameters and performance of the receiver. Section 6 concludes this paper.

2. Chirp FSK MICS modulation

2.1. Mathematical background associated with LFM

Linear Frequency Modulation (LFM) signal or linear chirp modulation signal is considered as a form of spread spectrum technique for wireless communications. Since 1940 chirp modulation has been widely used in radars communication systems because it has the property of pulse compression [9].

LFM has been chosen because it presents performances in terms of distortion, rejection of interference and allows a low power implementation [10]. Linear Frequency Modulation (LFM) signal or linear chirp signal can be representation by (1):

$$S(t) = A.\cos\left(2\pi f_0.t + \varphi(t)\right) \qquad 0 \le t \le T \tag{1}$$

Where f_0 the initial frequency of the chirp,

T is the duration of a single chirp pulse and the instantaneous phase $\left(2\pi f_0.t + \varphi(t)\right)$ varies linearly with time.

The instantaneous frequency $f(t)$ is given by (2):

$$f(t) = \frac{1}{2\pi}\frac{d}{dt}\left(2\pi f_0.t + \varphi(t)\right) = f_0 + \frac{1}{2\pi}\left(\frac{d}{dt}\varphi(t)\right) \tag{2}$$

If we make $\varphi(t) = 2\pi C_r.t^2 + \varphi_0$ where C_r and φ_0 are constants then $f(t)$ becomes (3):

$$f(t) = f_0 + 2C_r.t = f_0 + R.t \tag{3}$$

We can notice that the instantaneous frequency is a linear function of time with a slope R given by (4):

$$R = 2C_r = \frac{B}{T} \tag{4}$$

where B is the bandwidth. R is called chirp rate. Therefore the instantaneous frequency is increasing or decreasing over the chirp pulse duration and the bandwidth B equals the swept frequency.

For increasing instantaneous frequency (up chirp) the LFM signal is given by (5):

$$S(t) = A.cos(2\pi f_0.t + \pi R.t^2 + \varphi_0). \tag{5}$$

For decreasing instantaneous frequency (down chirp) the LFM signal can be represented by (6):

$$S(t) = A.cos(2\pi f_0.t - \pi R.(t^2 - 2T.t) + \varphi_0) \tag{6}$$

As represented in Figure 2, the transmission of a bit 1 corresponds to the band B1 (402.9 MHz - 403.2 MHz) and the transmission of a bit 0 corresponds to the band B0 (403.2 MHz - 403.5 MHz).

2.2. Transceiver architecture

The super heterodyne architecture presents problems associated with the use of several frequency transposition stages, such as the image frequency that must be taken into account during the design of the receiver. Moreover, this architecture needs an external IF filter which is a major problem to full integration.

In order to avoid image interference, the choice of the receiver architecture was the direct conversion (or zero IF) receiver. In zero IF architecture we can ensure the hardware minimization because it eliminates the image–reject filter and other IF components. Except Low Noise Amplifier (LNA) all amplification and filtering will take place at very low frequencies. Therefore, it is suited to a monolithic integration.

A Linear Frequency Modulation with Binary FSK scheme is proposed for this sensor node. This modulation has advantages over other techniques such as OOK and PPM. The reasons are: the demodulation operation are simple because it demands a zero threshold detector for the decision circuit and it is less affected by inter-pulse interference (IPI) for the same pulse rate [11].

The only disadvantage is that quadrature local oscillator waveforms are produced at the RF frequency. In consequence, the

power consumed will increase. But the LO frequency is about 403.5 MHz, which is considered relatively low frequency so the power consumption of the quadrature local oscillator is acceptable [12].

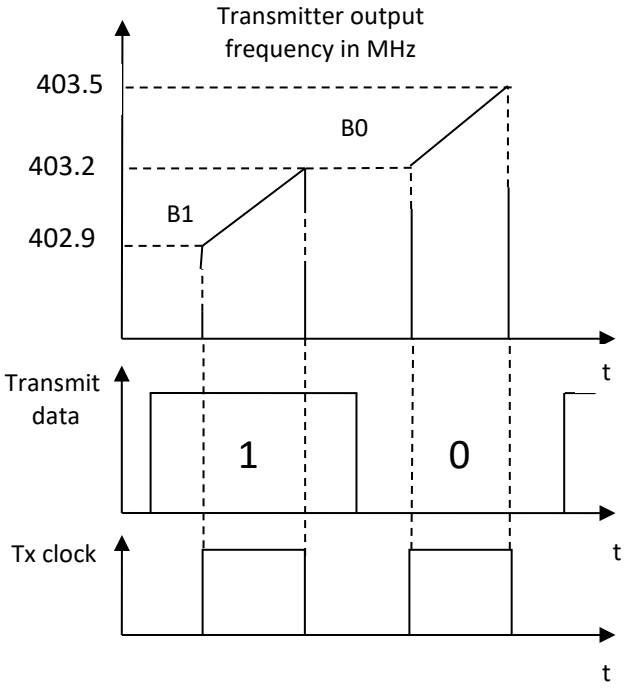


Figure 2. Output frequency of the transmitter in time domain

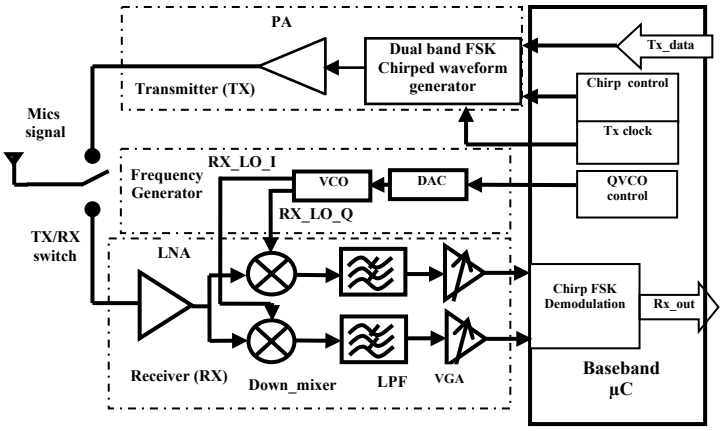


Figure 3. Illustration of zero-IF transceiver architecture.

Figure 3 shows different blocs of the proposed transceiver. The transmitter (TX) is composed of a dual band FSK chirped waveform generator and a power amplifier (PA).

A chirp control and Tx clock signals are used to switch between B0 and B1 signals depending on transmitted binary information and produces the LFM Binary FSK waveform. The power amplifier brings the signal power to the desired level in order to drive the antenna.

The RF band pass filter is not present in this design because it increases the noise figure and the circuit complexity of the receiver. In the other hand band pass filtering supplied by both antenna and LNA can mitigate the out-of-band signals at the receiver input.

After the antenna a LNA is used to amplify and band pass filter the received signal then the mixer down-converts it using quadrature LO signals. LO tones are produced with a VCO controlled by a Digital to Analog Converter (DAC). The DAC must generate the different voltages so that the VCO will be able to scan the entire RF frequency range from 402 MHz to 405 MHz with steps of 30 KHz. This technique is very useful for improving sensitivity and specially in presence of interferer. However, the system varies the LO frequency with a certain step to sweep the entire RF band and look for a clear channel and then perform the reception and demodulation. Before the VGA a low pass select filter (LPF) passes the desired channel and removes the other channels. Then a dual band chirped binary FSK demodulator based non coherent technique will recover the binary transmitted data [13].

3. Receiver design and simulation

In this work ADS tool is used for simulation and the receiver schematic is presented in figure 4. The aim of this simulation is to determine parameters for LNA, mixer, LPF and VGA that meet MICS and receiver specifications.

For this design the Pulse Rate Frequency (PRF) is set to 100 KHz and the bandwidth B is set to 300 KHz.

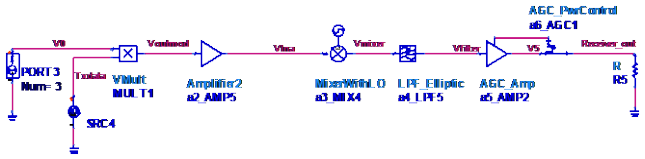


Figure 4. Receiver schematic.

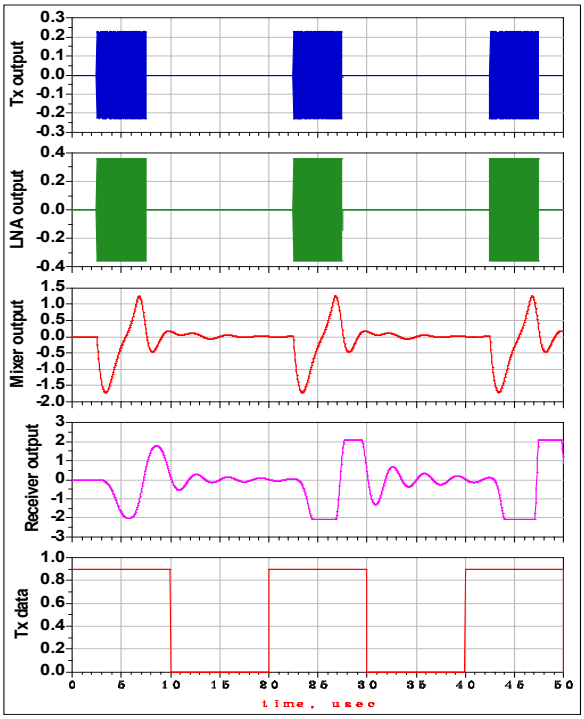


Figure 5. Tx binary data, Tx output, LNA output, mixer output and receiver output in time domain for t_p=5µs, Pin = -16 dBm and a PRF of 100KHz

3.1. Spectrum at each node of the receiver

A harmonic balance simulation was made for both minimum and maximum receivers input power in order to evaluate receiver performances. Figure 5 presents time domain signals of the transmitted binary data (Tx data), transmitter output (Tx output) as spread spectrum BFSK signal, LNA output, mixer output and receiver output for t_p=5µs, Pin = -16 dBm and a PRF of 100KHz.

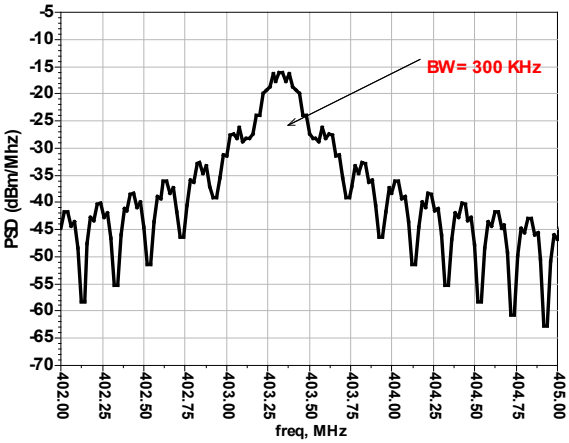


Figure 6. Receiver input in frequency domain

Figure 6 presents the spectrum at receiver input for channel number 5. We can see that the bandwidth B=300 KHz and the maximum input power Pin=-16 dBm. Figure 7 shows LNA output spectrum. Mixer output, LPF output and receiver output spectrums are shown in Figure 8 for maximum input power Pin=-16 dBm.

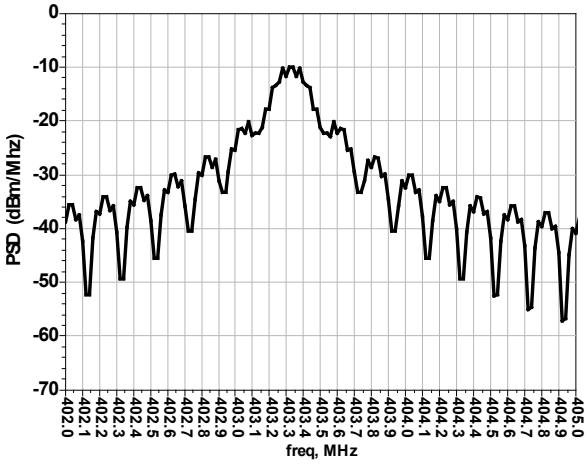


Figure 7. LNA output in frequency domain (Pin=-16 dBm)

It's clear that the result of simulation of the output Power Spectral Density (PSD) aligns FCC mask requirements such as bandwidth 300 KHz. Therefore, the receiver down converts the received signal properly.

Figure 9 presents receiver input in frequency domain for input power Pin=-93 dBm which is the receiver sensitivity. In the same way, Mixer output, LPF output and receiver output spectrums are shown in Figures 10 and 11 for input power Pin=-93 dBm.

From Figures 7 and 11 we can notice that the receiver output dynamic range is between -62 dBm/MHz and 0 dBm/MHz. This power level can drive the chirped Binary FSK demodulator stage.

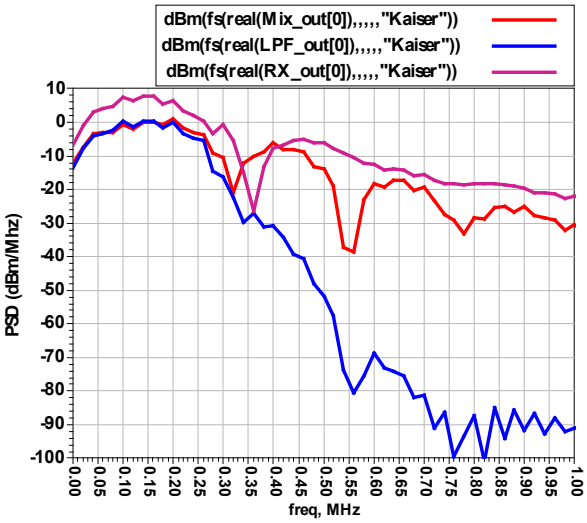


Figure 8. PSD for mixer, LPF and RX outputs, Pin=-16 dBm

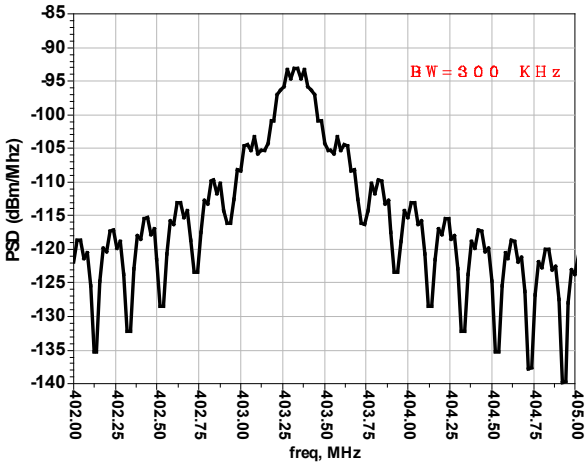


Figure 9. RX input in frequency domain for channel 5, Pin=-93 dBm

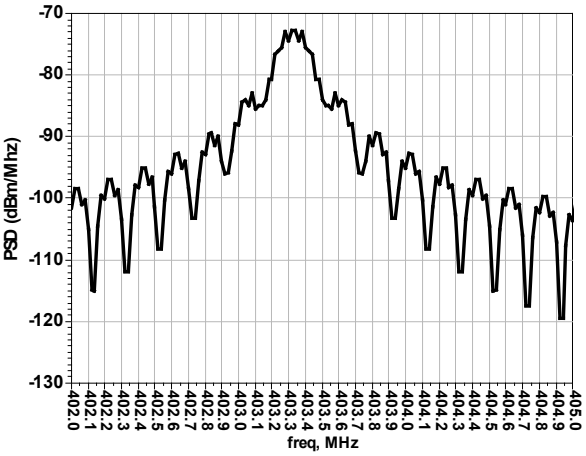


Figure 10. LNA output in frequency domain (Pin=-93 dBm)

3.2. LPF design

A variable bandwidth and sharp roll-off LPF has been used to select the clear channel from the 10 channels used in the MICS standard and reject at maximum interference. In this design we used a 5th order Elliptic filter with roll-off, 2dB rejection loss and

65dB attenuation. A tuning range from 40 KHz to 300 KHz is adopted as shown in figure 12. In addition this variable bandwidth Low Pass Filter can reject any narrowband interference by sweeping its bandwidth. This will improve the receiver performance.

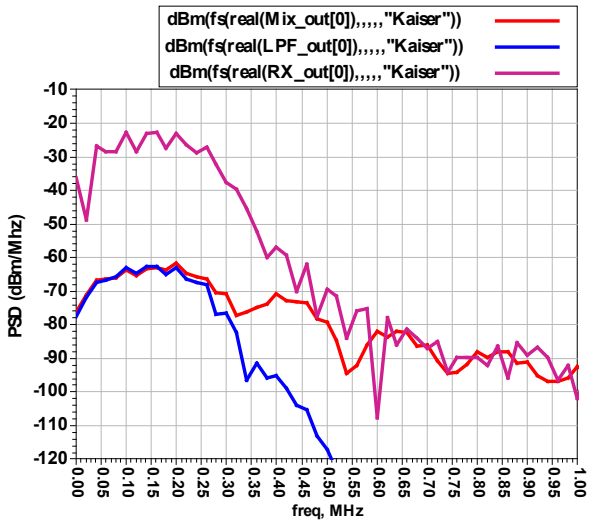


Figure 11. Spectrum for mixer, LPF and RX outputs, at a sensitivity Pin=-93 dBm.

3.3. Budget simulation

A Budget simulation for Pin= -16dBm was conducted which led to the following results: In Figure 13(a) the cascaded gain is 29dB, in Figure 13(b) the cascaded output power is equal to 13dBm and in Figure 13(c) the Noise Figure (NF) is 5.5dB.

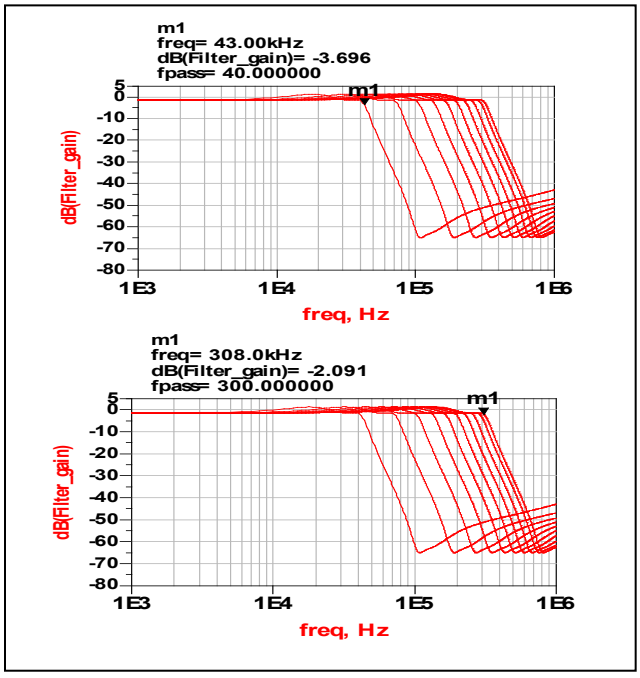


Figure 12. LPF response in frequency domain

For Pin= -93dBm (sensitivity) a Budget simulation was conducted which led to the following results: In Figure 14(a) the cascaded gain is 102dB, in Figure 14(b) the cascaded output power is equal to 10dBm and in Figure 14(c) the Noise Figure (NF) is 5.5dB.

Figure 13(d) demonstrates that the receiver achieves a P_{-1dB} of -22.1dBm and an IIP3 of -12.6dBm. This means that the receiver is linear up to -22.1dBm. The sensitivity of the receiver (-93dBm) is an excellent value compared with Bluetooth or Zigbee communication systems. Therefore this receiver can detect strongly attenuated input signals. A receiver NF of 5.5dB is acceptable to achieve a bit error rate of 10^{-3}.

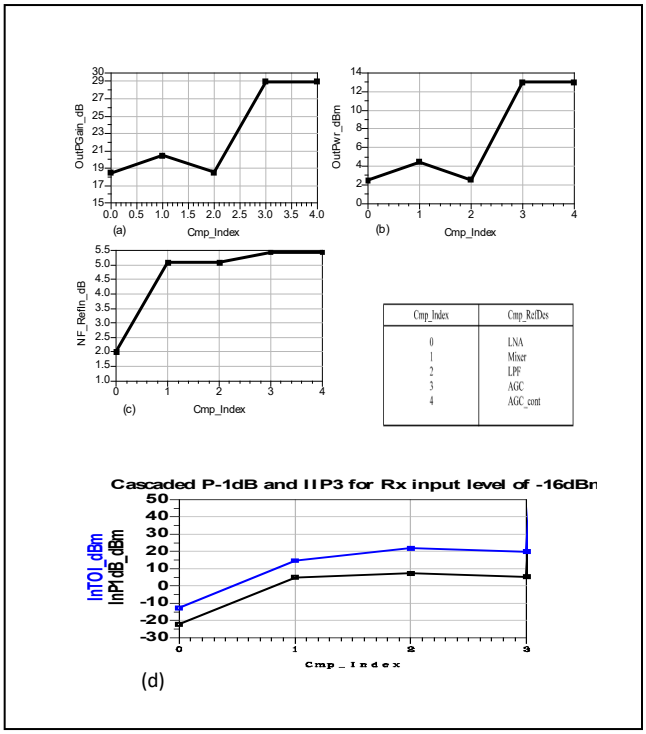


Figure 13. (a) receiver gain, (b) receiver output power, (c) receiver NF, (d) 1dB compression point and IIP3, for Pin=-16dBm.

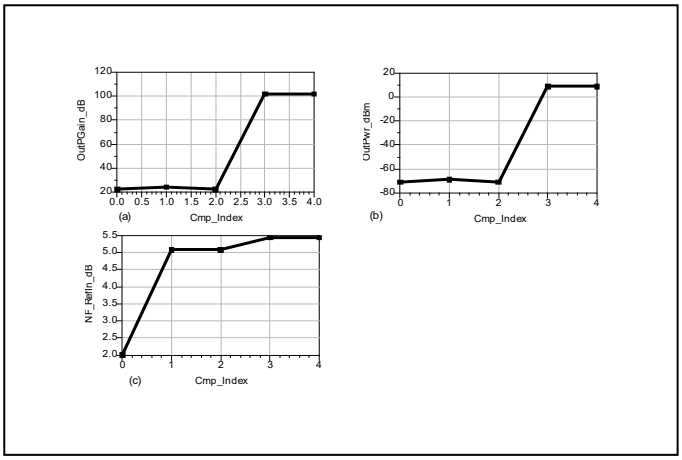


Figure 14. (a) receiver gain, (b) receiver output power, (c) receiver NF, for Pin=-93dBm.

4. FSK dual band Demodulator

The design of demodulator circuit for Binary Frequency Shift Keying (BFSK) modulation request simple circuits that consume low power. Non coherent demodulator based on an envelope detector, integrate and dump, a sampling and hold and a limiting

circuit still provides acceptable performance for low complexity MICS receivers, especially in multipath environments.

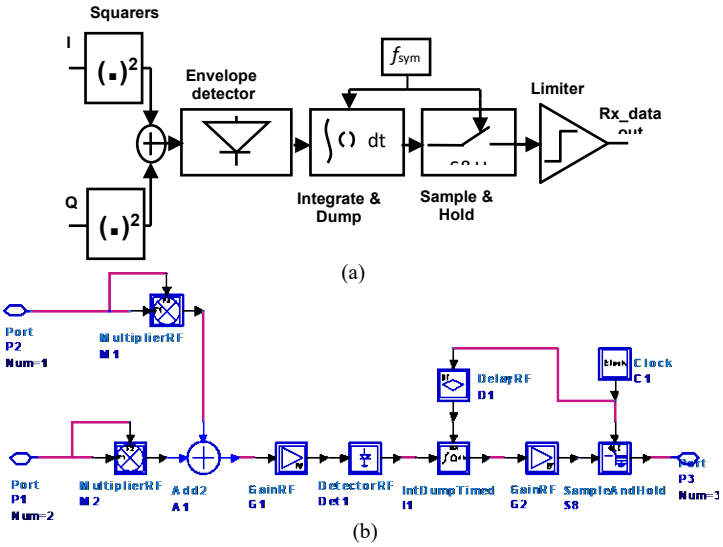


Figure 15. (a) block diagram of the Binary dual band FSK demodulator, (b) Implemented on ADS

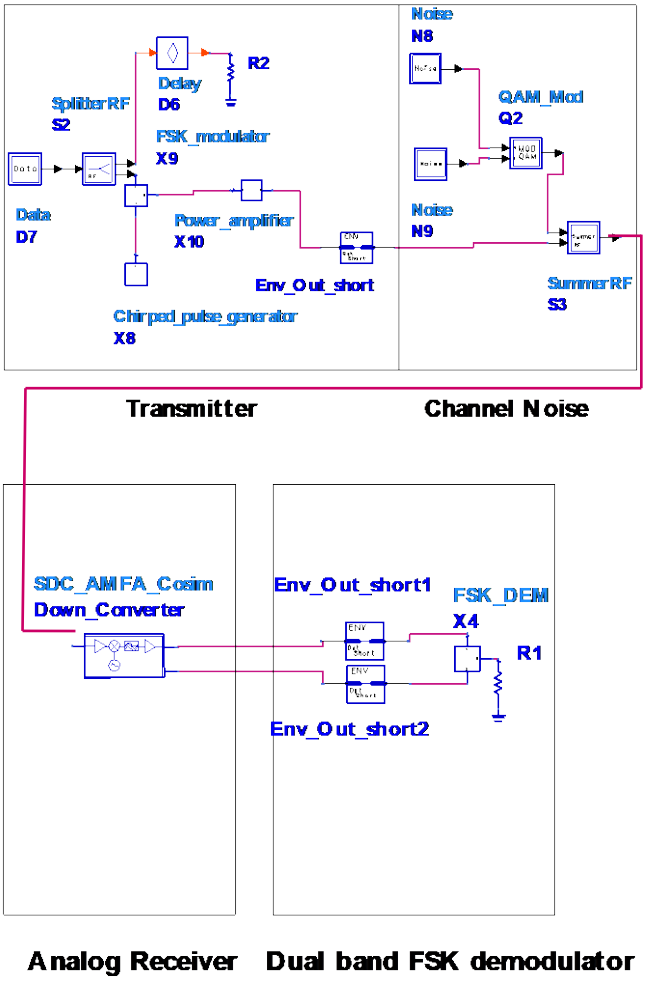


Figure 16. The template used to Co simulate the receiver with the chirped BFSK demodulator

Figure 15(a) presents the proposed FSK dual band demodulator. Advanced Design System tool was used to simulate this demodulator and the corresponding schematic is shown in Figure 15(b). Tow squarer (multiplier RF) are employed to perform energy detection. The I and Q signals are added (Add2). The combined signal is fed into an envelope detector then into integrate and dump block. After sampling-and-hold, a limiting amplifier will retrieve transmitted data. This technique provides acceptable performance for low complexity receivers, especially in multipath environment.

The analog transceiver with the chirped dual band FSK demodulator are co simulated and the corresponding template is shown in Figure 16. AWGN channel noise was used to run this co-simulation.

Figure 17 shows time representation of Tx data and chirped pulses for bit 1 and bit 0.

Figure 18 shows Tx spectrum for bit B1, bit B0 and the dual band Tx output spectrum. We can notice that the Tx output spectrum respects the FCC mask requirements for the MICS band. The emission power is under -16 dBm in order to avoid interferences with other communication services using the same MICS RF band.

The receiver output spectrum after low pass filtering and amplification is represented in Figure 19. The in band power of 0 dBm can drive the dual band FSK demodulator input.

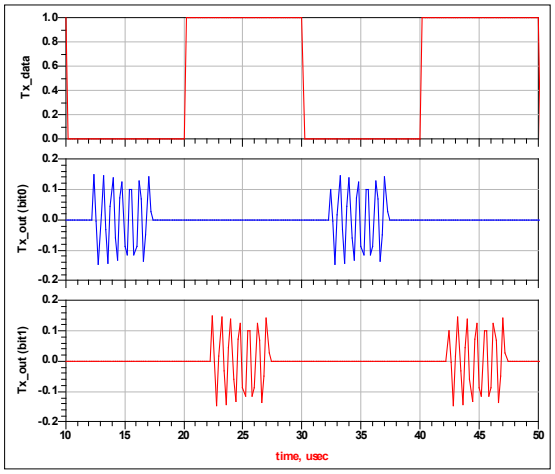


Figure 17. Time representation of Tx data and chirped pulses for bit 1 and bit 0

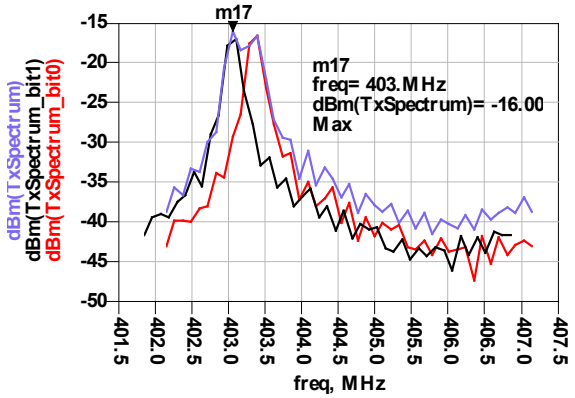


Figure 18. Tx spectrum for t_p=5µs and PRF=100KHz

Figure 20 shows waveforms at the adder output (adder_out), on the output of the envelope detector (env_det_out), after the block of integration (int_out), at the sample and hold output (S&H_out) and at the system output (Rx_dataout).

The co-simulation results show that the binary signal at the output of the demodulator is exactly the same as the delayed binary signal transmitted. Thus proposed demodulator is working properly.

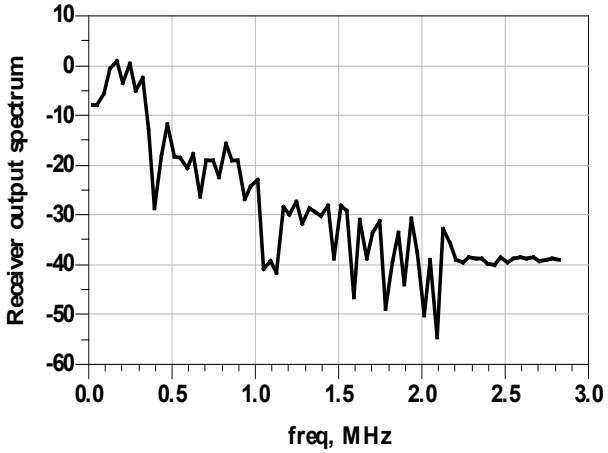


Figure 19. Output spectrum of the direct conversion receiver

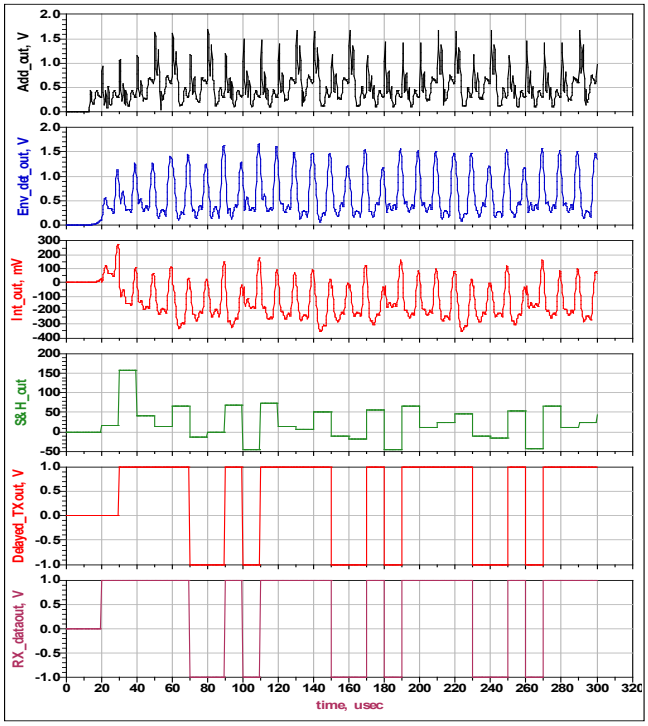


Figure 20. Time domain waveforms at different demodulator nodes

Table 1. Summary of the main technical parameters of the receiver

Receiver performance		**Receiver blocks requirements**			
Chirp FSK 402-405 MHz	**Bandwidth 300 kHz**	***LNA***	***Mixer***	***Var-BW-LPF***	***VG-Amplifier***
Noise figure	5.5 dB	S21= 22 dB	Con gain= 7 dB	Fpass= 40-300 kHz	Max_con gain= 80 dB
Gain	102 dB	S11= -13 dB	S11= -10 dB	Order N= 5	S11= -15 dB
IIP3	-12.6 dBm	NF= 2 dB	S22= -10 dB	Max rej= 65 dB	S22= -20 dB
Sensitivity	-93 dBm	P_{-1dB} = -22dBm	S33= -15 dB	ILoss= 2dB	NF= 13 dB
Data rate	100 Kbps	IIP3= -7dBm	NF= 7 dB	-	IIP3= 26 dBm

Table 2. Comparison with previously reported receiver

	This work	**[12]**		**[13]**	**[14]**	**[15]**	**[16]**
Frequency	402-405 MHz	402-405 MHz		3-5 GHz	402-405 MHz	433 /868 MHz	402-405 MHz
Modulation	LFM FSK	DBPSK	DQPSK	Dual Band FSK	FSK	FSK /OOK	BFSK
Bandwidth	300 KHz	300 KHz	300 KHz	500 MHz	300 KHz	-	300 KHz
Data Rate	100 Kbps	75.9 /151.8 Kbps	303.6 Kbps	3.2 Mbps	800 Kbps	100 Kbps	100 Kbps
Tx power	<-16 dBm	<-16 dBm	<-16 dBm	<-14 dBm	<-16 dBm	<-16 dBm	<-16 dBm
Max pulse width	5 µs	-	-	75 ns	-	-	-
Max PSD	-16 dBm	-	-	-41.3 dBm	-4.5 dBm	10 dBm	0 dBm
Sensitivity	-93 dBm	-92 dBm	-89 dBm	-	-83 dBm	-111 dBm	-70 dBm
Noise Figure	5.5 dB	6dB	6dB	5.32 dB	-	-	11-40 dB
Gain	102 dB	-	-	99 dB	-	-	35-51 dB
IIP3	-12.6 dBm	-	-	-8.8 dBm	-	-	-

5. Circuit blocks parameters and performance

The harmonic balance simulations results prove that the proposed design respects the constraints imposed by the FCC spectral mask and requirements of MICS standard.

In addition, the performance of the receiver is improved because it is able to reject any narrowband interference by sweeping the bandwidth of the LPF.

This chirped dual band BFSK demodulator can reply for the main challenge that is the low power. So, it is the best candidate for this kind of receiver. With this demodulator we can reach a better sensitivity.

In Table 1 we give a summary of the main technical parameters of the receiver showing its performance. The optimized specifications of the functions constituting the receiver are listed as well.

Table 2 shows a performance summary of the chirped BFSK MICS receiver. A comparison with other works dedicated for healthcare devices are listed as well.

It is proved that the proposed chirped BFSK MICS receiver designed using LFM technique and non-coherent dual band FSK demodulation has a sensitivity and a NF better than those found in the works [12], [14] and [16]. The results found in this design show that the proposed MICS receiver can be realized by simple, inexpensive and low consumption circuits which authorize healthcare implantable devices application.

6. Conclusion

A low-complexity LFM dual-band Binary FSK receiver for medical implantable wireless devices has been presented. A direct conversion architecture without RF band pass filter was used to avoid image interference and reduce circuit complexity. Linear Frequency Modulation technique was used to generate chirped Binary FSK signals. In order to reach low power implementation a simple non coherent demodulator was adopted. It is based on an envelope detector and simple decision circuits. Time domain, frequency domain and budget simulations results has been presented. We demonstrate that the proposed design respects the constraints imposed by the FCC and MICS requirements. A co-simulation of the analog receiver with the proposed non coherent demodulator shows that the receiver can reach a sensitivity of -93 dBm. In addition, the receiver is robust because it is able to reject any narrowband interference by sweeping the bandwidth of the LPF. This design allows the total integration of the sensor node in the same chip using low power CMOS technology. It promotes its use on a large scale in sensor networks implanted in the human body. This will essentially help people with disabilities, people with chronic diseases, people living alone and aged people to overcome their disabilities and live like healthy people.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Bettaieb, H. Trabelsi and M. Masmoudi, “A Novel MICS Chirp FSK Receiver Front-End” in International Conference on Engineering & MIS ICEMIS2017, Monastir, Tunisia, May 2017.

[2] Voskerician G., Shive M.S., Shawgo R.S., Von Recum H., Anderson J.M., Cima M.J., Langer R. “Biocompatibility and biofouling of MEMS drug delivery devices” Biomaterials, 24(11), 2003, pp.1959-1967.

[3] World Population Ageing 2015, Departement of economic and social affairs Population Division.United States.New York, 2015.

[4] H.S. Savci,A. Sula,E. Arvas, “MICS Tranceivers: Regulatory standard and applications”, in Proc.of Southest Con,April 2005, pp. 179-182.

[5] J. Ryckaert, M. Badaroglu, V. D. Heyn, G. V. der Plas, P. Nuzzo, A. Baschirotto, S. D’Amico, C. Desset, H. Suys, M. Libois, B. V. Poucke, P. Wambacq, and B. Gyselinckx, “A 16 mA UWB 3-to-5 GHz 20 M pulses/s quadrature analog correlation receiver in 0.18 µm CMOS,” in IEEE ISSCC Dig. Tech. Papers, pp. 368–377, 2006.

[6] FCC Rules and Regulations, “MICS Band Plan” , Part 95, jan 2003.

[7] D. C. Daly and A. P. Chandrakasan, “An energy efficient OOK transceiver for wireless sensor networks,” IEEE J. Solid State Circuits, vol. 42, no. 5, pp. 1003–1011, 2007.

[8] J. Ayers, N. Panitantum, K.Mayaram, and T. S. Fiez, “A 2.4 GHz wireless transceiver with 0.95 nJ/b link energy for multi-hop battery free wireless sensor networks,” in Symp. VLSI Circuits Dig. Tech. Papers, pp. 29–30, 2010.

[9] H. Trabelsi, Gh. Bouzid, F. Derbel and M. Masmoudi, “A 863-870MHz Spread Spectrum FSK Transceiver Design for Wireless Sensor,” Microelectronics Journal (MEJ), vol. 41, No 8, pp 465-473, August 2010.

[10] Simanjuntak, Lastri, “A Novel Chirp Slope Keying Modulation Scheme for Underwater Communication”, University of New Orleans Theses and Dissertations. Paper 201, December 2004.

[11] R. Roovers, D. M. W. Leenaerts, J. Bergervoet, K. S. Harish, R. C. H. van de Beek, G. van der Weide, H. Waite, Y. Zhang, S. Aggarwal, and C. Razzell, “An interference-robust receiver for ultra-wideband radio in SiGe BiCMOS technology,” IEEE J. Solid-State Circuits, vol. 40, pp. 2563–2572, 2005.

[12] BA et al, “ a 0.33nJ/bit IEEE802.15.6/Proprietary MICS/ISM Wireless Transceiver With Scalable Data Rate for Medical Implantable Applications”, IEEE Journal of Biomedical and Health Informatics, Vol.19 .No.3, May 2015.

[13] H. Trabelsi, I. Barraj and M. Masmoudi “A 3-5 GHz FSK-UWB Transmitter For Wireless Personal Healthcare Applications”; International Journal of Electronics and Communications (IJEC), Vol. 69, pp. 262-273, 2015, DOI:10.1016/j.aeue.2014.09.009.

[14] P. D. Bradley, “An Ultra-Low power, high performance medical implant communication system (MICS) transceiver for implantable devices,” in Proc. IEEE Biomed. Circuit Syst. Conf., pp .158-161, December 2006.

[15] V. Peiris, C. Arm, S. Bories, S. Cverveny, F. Giroud, P.Graber, S. Gyger, E. Le Roux, T. Melly, M. Moser, O. Nys, F. Pengg, P.-D. Pfister, N. Raemy, A. Ribordy, P.-F.Ruedi, D. Ruffieux, L. Sumanen, S. Todeschin, and P. Volet, “ A 1V 433/868MHz 25kbps-FSK 2kbps-OOK RF transceiver SoC in standard digital 0.18µm CMOS,” in Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers, pp. 258-259, February 2005.

[16] S. Kim, W. Lepkowski, S. J. Wilk, T. J. Thornton and B. Bakkaloglu, “A Low-power CMOS BFSK Transceiver for Health Monitoring Systems” in *IEEE Biomed Circuits Syst Conf.* 2011, 157–160. doi:10.1109/BioCAS.2011.6107751.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 67-79 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Revealing Strengths, Weaknesses and Prospects of Intelligent Collaborative e-Learning Systems

Amal Asselman*, Azeddine Nasseh, Souhaib Aammou

LIROSA, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, 93 000, MOROCCO

A R T I C L E I N F O

Article history:
Received: 09 April, 2018
Accepted: 02 May, 2018
Online: 20 May, 2018

Keywords:
Artificial intelligence
Collaborative e-Learning
Adaptive hypermedia techniques
Cognitive user model
Multi-agent Systems
E-Learning systems

A B S T R A C T

The rapid evolution of Collaborative e-Learning Systems migrates to the use of new technologies such the artificial intelligence (AI). In this context, the role of AI in increasing the quality of learning and making it more productive, persistent and efficient. In addition, it can accomplish repetitive and complex tasks in record time and unmatched accuracy. These advantages offer the ability to interact with learners in an almost human way. This interaction could be made on the base of adaptive hypermedia techniques, Multi-agent Systems technology and a cognitive learner model.

In this paper, we present and analyze some existing intelligent collaborative e-Learning systems on the basis of their various features such as collaboration features, intelligent actors' interaction, adaptability measurement, cognitive student modeling, and security measurement. Our analysis aims to provide important information to researchers, educators and software developers of educational environments concerning strengths and weaknesses of those e-Learning systems. According to this study, we found that some collaborative e-Learning environments, even the use of the mentioned technologies, still poor in terms of the structure of human cognitive architecture aspects and the capacity to assess the help provided to learners. For these reasons, we present, in the end, some prospects in order to determine how we can improve these systems to stop the reasons of abandoning courses.

1. Introduction

The evolution of the Internet and the development of educational content have led to the emergence of a new mode of education called e-Learning (electronic learning). This mode is used in several areas including employee training follow-up or self-training. E-Learning tends to develop around the world following the evolution of new Information and Communications Technologies (ICT) in the educational field that allows him to have many advantages such as the ability to facilitate teaching, and the accessibility of educational resources. This mode of learning is based on remote access to the provision of educational resources and learning services. The online learning environments are recognized by a whole different set of appellation vary according to the tasks assigned to them. Those environments remain in constant evolution, we cite WBT (Web-Based Training), LMS (Learning Management System), LCMS (Learning Content Management System) CLCS (Computer Support for Collaborative Learning), MOOC (Massive Open Online Course) etc.

Generally, the analysis of e-Learning systems based on their characteristics and functionalities [1] [2] mention that each of them is credited for enhancing the acquisition of learners. They offer them the opportunity to learn using a modern, fun, and interactive strategies that are completely different to traditional learning in which requires the memorization of a set of knowledge to better get them back on the day of examination. These programs also gives to the learner the choice of lessons that meet his needs without obligation to follow the lessons imposed by the teacher.

Despite all their benefits, these systems have recognized a major obstacle due to poor management of learning time and lack of orientation on the platform. In addition, the geographical dimension causes a difficult finding of available human accompanist in e-Learning environment during the whole day and throughout the learning period. Moreover the difficulty of finding always learners with the same profile available to collaborate. As a result, the learner can simply be lost depending on the huge number of resources and links. In fact, this eventually leads him to

*Corresponding Author: Amal ASSELMAN, LIROSA, Faculty of Sciences, Abdelmalek Essaadi University, Tetouan, 93 000, MOROCCO
E-mail : asselman.amal@gmail.com

feel that online training is boring and decide automatically to drop out.

According to the reason mentioned before, several statistics [3] [4] [5] [6] have been realized showing us the number of students training leave, as example Figure 1 concerning "MITx course, 6.002x (Circuits and Electronics)" available on edX online learning platform [7]. This example indicates the number of learners participated during all phases of the training. It illustrates that there is a high drop-out rate appears from the first week. And after completing the 14 weeks of study, it remains a few of student arrive to accomplish this course. In terms of numbers, there are approximately just 4.6% of students have succeeded while the 95.4% withdrew.

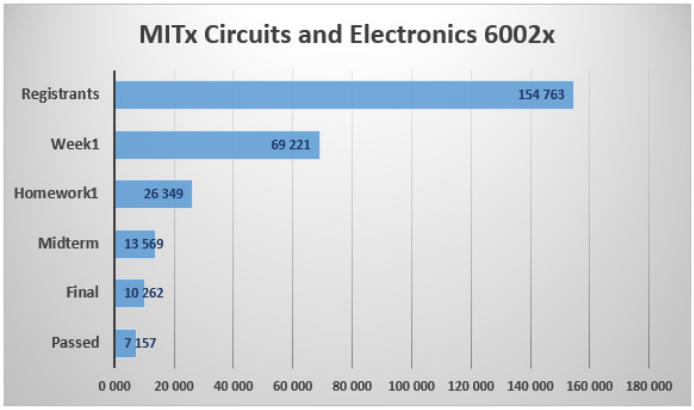


Figure 1: Dropout rate among those who initially registered [7]

Consistent with this context and precisely with the accelerated development of information technologies as well as the trend towards the use of smart devices, the integration of artificial intelligence makes a great change in the society in the context of Collaborative e-Learning Systems. It remains as an appropriate solution to make educational system more productive, persistent and efficient [8]. It allows the creation and extension of various intelligent technologies such as adaptive hypermedia techniques, agent technology, and web services which makes e-Learning more effective and flexible. In the same case, the use of Multi-agent Systems (MAS) technology in the design and modeling process of learning environments has evolved spectacularly. It proved its ability in modeling the different actors of e-Learning systems and managing their interactions to meet their dynamic and execution needs by their innovative features like intelligence, autonomy, reactivity, social ability, and proactivity. However, even though there are collaborative e-Learning environments using the mentioned technologies, they have structured with a minimum of consideration the relevant aspects of human cognitive architecture, a low capacity to assess the help provided by learners and several other weaknesses.

This paper, as an extension of work originally presented in 15th International Conference on Emerging eLearning Technologies and Applications (ICETA) [9], is organized as follows. Section 2 presents general principles of collaborative online learning by presenting the different activities of e-Learning system's actors as well as the interaction between them. Section 3 presents an overview of distributed artificial intelligence, multi-agent systems, and adaptive systems in order to indicate their importance for improving the educational systems. Section 4 is based on an analysis presenting the advantages, the limitations of some existing agent-oriented collaborative learning system, and comparison between them on the basis of a multitude of criteria: collaboration features, intelligent actors interaction, adaptability measurement, cognitive student modeling, and security measurement. In the last, we will present certain recommendations in order to indicate how can improve these systems for stopping the main reason for drop-out.

2. Collaborative e-Learning Systems

2.1. Features and advantage of e-Learning

E-learning becomes an essential part of modern education that allows learners to consult the pedagogical content effectively everywhere and any time [2]. These contents are organized into modules that can be assembled into personalized training courses. Currently, many studies show that e-Learning training has better results compared to traditional classroom training. It is considered as an efficient and attractive way for computer education delivered or mediated a pedagogical content and learning experiences by means of digital technologies. These systems are recognized as a software environment designed, specifically, to meet the needs of distance education and to allow organizing the online content in which the authors have a simple interface to deposit their contents (within predefined themes and organized in categories). They can associate one or several downloadable files to their courses [2]. These systems are characterized by a number of advantages like:

- The ease of sharing, exchanging and using varied learning modalities (audio sequences, videos, animations, diagrams);
- Flexibility and adaptability according to learner availability (By providing training at any time or by allowing the learner to learn at his/her own pace and in an individualized way);
- The relative reduction of costs for learners and trainers (elimination of accommodation and travel costs) and increased accessibility of training.

E-Learning follows a new economic model of production in which a large number of people join with new communication tools. Learning is an active process where tutor plays a leading role in training to help advance learners by placing more emphasis on monitoring, accompanying rather than transferring expertise. The e-tutor occupies several goals, for example, he animates the group or the community of learners, he determines the most adapted training course to learners, he offers him the help in his first steps and he ensures the pedagogical follow-up of the training (personalized advice, analysis of the progression, answers to the questions of the learners). As in traditional learning systems, the teacher was previously the sole holder of knowledge. However, in modern e-Learning systems, his role is to facilitate learning by guiding learners to make good use of knowledge [10]. He can also observe the progress of student's activities in order to motivate and guide them during their prevent failure or abandonment. Moreover, the role of the learner is no longer similar to his / her role in traditional learning which is limited to the memorization and evaluation of his knowledge on the day of the exam, but he governs his process learning by transforming information into

knowledge and creating interactions with other members of the group. Before learning, students set goals and learning tasks of the plan. During learning, they work together to accomplish tasks and monitor their progress. And after learning, they evaluate their performance and plan for future learning [1].

Online education systems revolve around a set of functionalities such as:

- Administrative management of the platform. It concerns the management of registration in training and the management of administrative data training.
- Creation of courses and training plans (creation of teaching scenarios, curricula, and preparation of the various evaluation tests) besides changing or add course content and activities, schedule management and training path planning.
- It allows classification, indexation, and administration of pedagogical materials. As well as consultation of educational content, individualization of learning and communication between trainers and learners and between learners.
- It determinate roles of trainers, groups of students and their access by specifying content types, communication modes, test types, and by facilitates communication with students or between them by messaging or forum.
- It authorizes creation and modification of individualized training paths as well as follow-up of the activities of learners on the platform (time spent online in the course notes, dialogues, work deposited on the platform...).

E-learning researchers aim to improve distance learning systems in two areas: organizationally and functionally. The first by taking note the standardization of self-training materials and a quality approach, providing training opportunities and broadening the prospects for dissemination to employees, suppliers, and customers [1]. The second by trying to personalize the learner's profile according to training, to reduce costs, to save time and to make learning more effective.

2.2. Definition and principles of collaborative learning

Since the evolution of the web, the semantic field of the word "collaboration" has evolved considerably. Several education experts point out that school is not just a place to learn facts. It is a place for a student to interact with one another and learn the basics of communication, etiquette, and respect for others. Actually, many studies have focused on the collaborative e-Learning system that is defined as the learning strategy where several learners interact with their partners, share ideas and help each other to achieve their common goals [11]. These educational environments are designed and dedicated to encouraging individual work and group work in various fields involving the domain of instruction. In the same context, Andrew Carnegie as an American Industrialist and philanthropist define collaborative work as follows:

"Teamwork is the ability to work together toward a common vision. The ability to direct individual accomplishment toward organizational objectives. It is the fuel that allows ordinary people to achieve extraordinary results".

In collaborative e-Learning, actors can work together regardless of geographic location or time limitation. In other words, it provides learners with a great flexibility of time and place as well as excellent asynchronous interaction [12]. This mode of learning follows a learner-centered approach by involving new roles for teachers, accompanists, and learners. It can be reported that students can undertake problems that are more complicated and gain a better understanding of the material by working collaboratively [13]. For that, learning content is becoming increasingly hypermedia intelligent and collaborative, which allows placing initiative and power in the hands of learners to access the information. Moreover, this system makes the responsibility of collective and global learning. Indeed all members of the group stay in regular contact, each learner must participate in the group in an action to be realized, each member can contribute to the action of other members of the group to increase its performance. According to Walckiers and De Praetere in the case of collaborative work, there is no a priori distribution of roles: individuals are gradually subsumed into a group that becomes an entity in its own right [14].

Generally, Collaborative e-Learning provides functionalities essential in the educational process, such as real-time, as well as offline data and information gathering, analysis and distribution, embedded feedback, assessment, and collaboration. All of these functionalities are based on synchronous and asynchronous learning tools for interaction [15]. Where synchronous tools allow real-time communication between people remotely geographically: instant messaging, voice telephony, audio conferencing and video conferencing, etc. Indeed, the actors must be at the same time facing their respective computers. While asynchronous tools allow for time-and-space exchanges are e-mail, forum discussions, portfolio, wiki, and blog. In this type of system when a person sends a message, the receiver can read it for a few minutes, hours or days later.

Accordingly, Collaborative learning is a construct based on several sources. It feeds on the values of constructivism and relies on cognitive theories to explain the mechanisms of learning. For this reason, it promotes the integration of students in homogeneous groups which have the same profile and the common cultural and social aspects. This integration allows the development of critical thinking in the learner, as well as the emotional and pedagogical support of weak students. That can create for learners a source of motivation and consequently increase the communication skills, it can also support the realization of a continuous formative evaluation.

3. Artificial Intelligent technologies For Education

John McCarthy, an American computer and cognitive scientist pioneer and inventor, was known as the father of Artificial Intelligence (AI). The term AI was proposed, in 1956, by this person in Dartmouth Artificial Intelligence conference [16]. It was the first artificial intelligence conference organized to draw the talent and expertise of others interested in machine intelligence for a month of brainstorming. It refers to the ability to interact with the user and respond to their actions in a natural way [17]. It is recognized as a computer discipline that aims to model or simulate so-called intelligent human behaviors such as perception, memorization, decision-making, understanding, and learning [16].

Among the areas where artificial intelligence is ready to make big changes is the field of education. The application of AI to eLearning content is not just a cost-saving solution; it also opens up a whole new way of looking at learning itself.

Several advanced research has been done in this area, particularly in the field of intelligent tutorial systems (ITS) so that research aims primarily at making learners benefit from technological advances in artificial intelligence. It can adapt educational software to student needs by putting more emphasis on certain subjects, repeating things that students haven't mastered, and generally help students to work at their own pace [17]. AI will shift the role of the teacher to that of a facilitator so that teachers will supplement AI lessons, assist students who are struggling, and provide human interaction and hands-on experiences for students. AI-driven programs can give students and educators helpful feedback by allowing students to get the support they need and choose majors based on the areas they are doing well. And by providing for teachers ability to find areas where they can improve teaching for students who may be struggling with the subject. Data powered by AI can change how schools find, teach, and support students as well as it can point out places where courses need to improve. It also may change where students learn, who teaches them, and how they acquire basic skills [18]. Similarly, it is altering how can find and interact with information. In like manner, AI can make trial-and-error learning less intimidating in view of the idea of not knowing the answer or even failing is crippling. For this, artificial intelligence could help students learn in a relatively non-judgmental environment, especially when AI trainers can offer improvement solutions.

So, the use of artificial intelligence remains as an appropriate solution to make educational system more productive, persistent and efficient [8]. In order to achieve this intelligence, it must be aware of the user's objectives and knowledge for the purpose of designing adaptive educational intelligent systems. These systems try to help users during their learning by limiting the browsing space, suggesting the most relevant links to follow, or providing adaptation comments to visible links [19]. Add to that, integration of multi-agent systems offer an original approach to designing intelligent and collaborative systems. Multi-agent systems are characterized by the distribution of global control of the system, and by the presence of autonomous agents evolving in a shared and dynamic environment. They are more adapted to designing a learning environment where each member must manage and exchange knowledge, and collaborate with others to achieve its goals [15]. The fundamental aspects of these intelligent e-Learning systems are the ability of perception, inference, learning, reasoning, and knowledge-based systems [20], resulting from the integration of the agents' cognitive architecture. This latter aims to treat the process of development and use of knowledge through the set of mechanisms of human cognition. In other words, it models specific tasks based on the simulation of human behavior to create a model able to understand, reason, and solve problems [21].

Due to the benefited of artificial intelligence, we will be interested in the modeling of intelligent behaviors that are the product of the cooperative activity between several agents in order to design an adaptive system respecting the personal needs of each learner and to have agents capable of interacting similarly to humans thanks to the integration of cognitive architecture.

3.1. Agent and Multi-Agent System in education

The design of a distance learning system involves a complex set of processes that are sometimes difficult to associate in a coherent and evolving system. Multi-agent systems (MAS) are proving to be relevant for solving these types of problems thanks to its great ability to structure knowledge transfer systems and open new perspectives of assistance to distance learning. MAS is currently considered as the most active research discipline which they relate to several areas, in particular of artificial intelligence, distributed computing systems, robotics and software engineering. These systems are made up of several flexible and autonomous entities interacting with each other and called agents. According to Ferber, a multi-agent system can be considered as a macro-system composed of autonomous agents that interact in a common environment in order to achieve a coherent collective activity [11].

In literature, the notion of agent remains relatively vague to define. It is found that several researchers have defined this concept in different ways. All of these definitions look the same, but they also differ according to the application context for which the agent is designed. According to Ferber, an agent is defined as a physical or virtual entity, autonomous, located in an environment, capable of acting in an environment, communicating with other agents, perceiving its environment, reproducing itself [11]. This definition is supported by Wooldridge [22] in which present an agent as computer system located in a certain environment capable of autonomously performing actions in that environment, achieve its goals. Another frequently cited definition is presented by Russel [23]; he provides that an agent is all that can be seen as perceiving its environment through sensors and acting on this environment through effectors. Generally, agents aim to reduce the complexity of problem-solving time by dividing it into sub-problems, each sub-problem is assigned to an independent intelligent agent called "resolver". To achieve this, agents are organized, negotiated and coordinated for the purpose of resolving a common goal [24].

Generally, Educational area based on multi-agent systems presents a series of advantages for reason that these systems have inherited the traditional benefits of distributed resolution and artificial intelligence. Among which we can mention that intelligent systems are open, i.e. they have the ability to dynamically add or remove features or services in the agent system. It recognized a flexibility of the computer tool which aims to make the programming simpler, to change agents' behavior, and to add or to delete possible actions. They also have the advantage of the decentralization of the agents which aims at allowing to support the individual failure of these elements as well as distributed problem solving where a problem can be broken down into subparts, each of which can be solved independently to arrive at a stable solution. Moreover the speed of execution where several agents can work at the same time for the resolution of a problem thanks to the parallelism of execution.

3.2. A required architecture for e-Learning system

In order to build an intelligent educational system that is able to minimize the intervention of a human in the execution of tasks, to facilitate localization and customization of appropriate e-learning resources, thus to promote collaboration in e-learning environments, as well as to be able to automate effectively learning

and administration tasks, and behave axially like a human, it is necessary to implement an architecture containing agents able to act similarly as human users. To achieve this goal, the characteristics of each type of agent must be specified.

Agents can be classified based on the technology used to implement it, on the type of the agent or on the application domain itself. According to Wooldridge, agents can often be categorized according to their individual behaviors. Generally, two main categories of agents can be distinguished [22]:

- **Reactive agents:** have not any representation of other agents or their own environment as well as their inability to account for its past actions. They behave in a "stimulus-response" way in the face of what they perceive. This kind of agents argues that it is not necessary for agents to be intelligent individually for the system to behave intelligently [11]. In general, this category of agents only has a reduced communication protocol, which shows its weak communication capacity [25].
- **Cognitive agents:** are most represented in artificial intelligence distributed field. They are composed of a set of agents "intelligent" that include memory, attention, communication, and learning. Each agent has a knowledge base comprising all information and know-how necessary to carry out its task and to manage interactions with other agents and with its environment [26]. These agents have a global representation of themselves, their environment and other agents with whom they communicate. In addition, they also have an explicit representation of beliefs, desires, and intentions.

This description shows that an agent is capable of acting autonomously according to the goals that it pursues. Thus it can be seen that cognitive agents are more beneficial than reactive agents. Where the latter impose more rigid behaviors, which shows that reactive agents are not very powerful since they are reduced to their own means [26]. Moreover, it does not have a complete representation of its environment and other agents. For that, we can summarize that the most appropriate architecture for an intelligent education system is the one that uses a set of human cognition mechanisms based on cognitive agents. And consequently, due to their performance, the integration of this type of agent allows the system implemented to benefit the use human knowledge process. Specifically, it is based on the simulation of human behavior to create a model capable of memorizing, learning, understanding, reasoning, and solving problems [27].

The e-Learning's cognitive architecture defines the manner in which a cognitive agent (learners, administrators, trainers ...) manages the resources that it has primitive. According to Anderson "A cognitive architecture is a specification of the structure of the brain at a level of abstraction that explains how it achieves the function of the mind" [28]. According to [27], we compared between a set of cognitive architecture and conducted a detailed functional comparison by looking at a wide range of cognitive components, including perception, goal representation, memory types, learning mechanism and problem-solving method. This comparison and other similar analyses [29] [30] [31], defined ACT-R as is the most appropriate architecture for e-learning systems which it allows to have a behavior more identical to that of a human compared to the other architectures. ACT-R is applicable with multiple domains. In Fig.2 we distinguish some of their application.

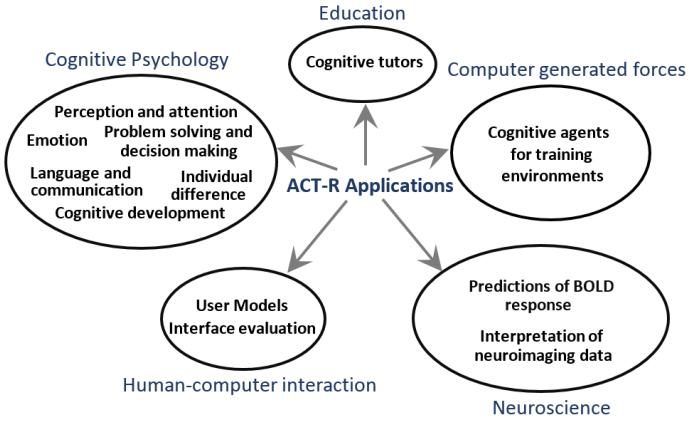


Figure 2: The main applications of cognitive architecture ACT-R

Recently, researchers on Cognitive science focus on the working process of the human brain that is resembled in various real-time applications. They show that the design of a learning system that is based on the integration of Anderson's ACT-R learning theory could be able to properly analyze a learner's behavior and know his or her cognitive state. It considerate each student has a cognitive profile and which can help the individual to be able to develop his or her learning skills and strategies in the light of useful self-knowledge. The system should learn like a human and should recognize the new skill, which is the main objective of cognitive science.

3.3. Adaptive learning

Adaptive hypermedia is a field that recognizes a strong use of artificial intelligence. The most popular hypermedia systems are related to the pedagogical field [32]. The hypermedia educational systems allow the creation of adaptive documents in order to establish some form of a dialogue between users and system. These documents stored on a computer support, as types of nodes connected by links [33]. Such systems offer to users the possibility to navigate in the hyperspace by allowing them to create their own educational path. The construction of the hypermedia is started when learner connects to visualize a course. In particular, the generation of content starts when the learner decides to activate a concept he wants to follow, or when he clicks on a hypertext link that leads him to another concept of the same course or another course [34].

In general, there are two kinds of hypermedia systems differ in the manner of adaptation: The adaptable hypermedia system and the adaptive hypermedia. The adaptable hypermedia offers to the learner the opportunities to configure the system by changing certain parameters, and consequently, this system can adapt its behavior [34]. The adaptive hypermedia system is characterized by its ability to update dynamically the user's profile during all phases of learning. This adaptation is done in different ways, either by using the user's navigation actions or by analyzing his answers to the questionnaires, it can be also based on the initial information provided by the user to adapt the nodes and navigation [34].

Moreover, the system can track and analyze the learner behavior and save it in a user model in order to adapt automatically the presentation consequently according to the characteristics of this model [35].

An educational hypermedia system is a system capable of adapting itself to the different characteristics of the students that are specified through a model. Brusilovsky [33] specifies that "Adaptive hypermedia systems are hypermedia systems which reflect some features of the user in a user model and use this model by adapting various visible aspects of a system to user". Adaptive hypermedia aims to improve education systems by enhancing the assimilation of knowledge through the adaptation of links, content, and presentations to the user. These systems can considerably reduce the user's path in hyperspace in order to avoid disorientation of user and the risk of misunderstanding the document.

An adaptive teaching system implements mechanisms that leverage on domain knowledge, learner knowledge, and knowledge about learning processes. This adaptive teaching offers approaches and personalized educational content. Usually, in order to increase the profitability of an educational application for a learner or a group of learners that have different profiles, it is obviously necessary to adapt the system according to learner goals, characteristics, and interests. This adaptation can be done in two ways: presentation adaptation and navigation adaptation. The first kind of adaptation makes it possible for each user to view a personalized page that is different from that of the others [11]. Where the navigation adaptation is interest in helping learners to find their self in hyperspace or to oblige them to use certain links rather than others in order to follow the most relevant path [19]. This phase involves changing the navigation structure or defining how this navigation structure will be presented to the user.

Learner modeling aims to give a complete and faithful description of all aspects of the user's behavior. It represents the core of adaptive learning system. When a learner takes training in a distance learning session, the learner model evolves according to learners' educational backgrounds and their responses [21]. This information can be directly entered by the learner (response to questions asked by the system), calculated from results of exercises or tests, or deduced from the learner's behavior in terms of interaction with the system (navigation choice, document reading time, etc.). This model aims to present all the characteristics of learners helping to achieve adaptation [36]. Those characteristics are separated in learner model into two categories [26]: Domain Dependent Data and Domain Independent Data. The first reflects the status and level of knowledge and skills which student achieved in a certain subject and at the certain moment of time. This category is organized by a set of elements (concept, topic, subject…). The second includes information about the learner skills based on his behavior. This category may include generic information (demographic information, learner learning goals, interests, background and experience) and psychological information such as memory, perception, learning style and preferences, decision making abilities, critical thinking, analytical thinking for modeling cognitive aptitudes, and motivation, reflection, self-awareness, self-assessment, self-monitoring, self-regulation for modeling metacognitive affective features…) [31].

4. Analysis of Agent-Oriented Collaborative Learning System

4.1. Existing intelligent collaborative e-Learning system

Many multi-agent systems that focus on the collaborative e-Learning domain have been designed and implemented by integrating a wide range of functionalities. These systems offer an original approach to designing intelligent systems and provide an interesting solution for both problems of structuring and exploiting knowledge. Agent-oriented collaborative learning system offers many appropriate tools for designing a learning environment where each member must manage and exchange knowledge and then to collaborate with others in order to achieve its goals. The common goal of these systems is to improve the learning ability of the student/learner., among which we can distinguish. In the field of scientific research, there are several systems that meet these objectives. In this paragraph, we will present and analyze them it in order to clarify their strengths and limitations.

4.1.1. ALLEGRO

ALLEGRO is an intelligent environment [37]. This system offers flexibility, autonomy, and adaptability to the e-Learning environment base on MAS. It offers an individualized learning as well as CSCL (Computer Supported Collaborative Learning) which makes the system capable of supporting collaborative learning. This system is based on three theories of learning behaviorism, cognitivism, and historic-social.

This system has six types of artificial agent:

- **Student Agent**: aims to manage the learning student model and maintains individualized information about the apprentice.
- **Tutor Agent**: will guide the learning process, decides what action to teach, how and when, It tries to detect mistakes in the process of the apprentice, in addition, it offers suggestions, critics, and recommendations through the selection of the suitable pedagogical strategy. This agent aims to promote learner learning by asking Agent Diagnosis to send to the apprentice an appropriate evaluation according to its profile. Or by asking Expert Agent to offer certain knowledge to the learner according to the plan and its diagnosis.
- **Diagnosis Agent**: is responsible for filtering and classifying the knowledge level of the learner.
- **Expert Agent**: manages the content of the area or specific subject of learning and teaching. It sends knowledge to the apprentice when he asks for it or at the request of the Tutoring Agent.
- **Collaboration Agent**: aims to group the apprentices by study topics, profiles or behavior and that's why it tries to find students who are interested in the same subject in order to create synchronous or asynchronous collaborative communication.
- **Interface Agent**: allows interaction between the user and the artificial agents. It allows the unfolding of knowledge and collaboration on the screen of the student.

4.1.2. BAGHERA

BAGHERA is a platform for distance learning [38]. This platform has been implemented with the JATLite multi-agent application development environment. It proposes the basics of computer environments for human learning (EIAH) seen as an educational community made up of human and non-human agents who cooperate and work together to train students. Its functional objective is to construct a logical experimental, flexible, and adaptable platform for distance education. Its high-level MAS is used for collecting what is needed for its pedagogical behavior, while its lover level MAS is responsible to diagnose student's conception. These diagnoses are based on the student's action towards the Interface

In the environment, each apprentice is supported by three types of agents:

- **Companion-Student Agent:** aims to make the link with the rest of the system and the students. Also through this agent, the student can access his electronic binder and a specific graphical interface that will allow him to work on one of his exercises, save his work and request a verification of the proof he develops.
- **Mediator Agent:** Chooses an appropriate agent to send the student's solutions for verification.
- **Tutor Agent:** He is an artificial designer. This agent is mandatory for every student registered in the system. When the teacher is absent, the tutor agent takes charge of the pedagogical follow-up for autonomous learning with regard to his student and the learning situation.

Similarly, the teacher is supported by two types of agents:

- **Companion-Teacher agent:** allows realizing the interface between the system and the teacher. It ensures the access of the teacher to his electronic locker. The teacher relies on the Companion-Teacher agent to connect with the students and other teachers who are available.
- **Assistant Agent:** The main function of this agent is to manage the teacher's record that contains all the exercises. The agent also manages the distribution of these exercises. For this purpose, he distributes them to the tutor agents who then transmit them at the appropriate time in the electronic binders of their students.

4.1.3. I-MINDS

Intelligent Multi-agent Infrastructure for Distributed Systems (IMINDS) [39]. It provides a multi-agent infrastructure for Computer-Supported Collaborative Learning (CSCL) that is capable of monitoring and tracking both students and teacher activities and making decisions to support the users. This system allows the learner to learn, interact with other learners as well as with the tutor and it also allows it to form a group. I-MINDS provides standard online collaborative features. The infrastructure is also capable of machine learning, allowing the agents to improve their performance over time or to adapt to individual user behaviors [39]. In this system, intelligent agents interact to serve tutors and learners. I-MINDS provide collaboration in two principle manner: forum and whiteboard. During Forum communication, student agents recognize some aspects of tracked behaviors including the number of messages each student contributed to the forum, the average length of each message, and the average quality of each message. Also during whiteboard collaboration, student agents track behaviors including the time that each student spent on the whiteboard and the tools (e.g., annotation, drawing, and eraser). This architecture provides to a teacher the ability to send their lectures directly from the classroom to the students via whiteboard technology. I-MINDS consists of three intelligent agents:

- **Student Agent:** manages the interaction channels among students and exchange information between the teacher and the students and the group agent. The student agent presents the learning material to the student. It can also assess and form a buddy group for the student that it serves. Other aspects of tracked behaviors include the number of messages each student contributed to the forum, the average length of each message, and the average quality of each message;
- **Group Agent:** is designed to encourage collaborative learning groups. It controls the students' interactions to assess each student's contribution as a group member.
- **Teacher Agent:** interacts with a teacher. It is responsible for disseminating information streams to student agents, maintaining profiles for all students, assessing the progress and participation of different students, ranking and filtering the questions asked by the students, and managing the progress of a classroom session.

4.1.4. MASCE

MASCE is the acronym of Multi-Agent System for Collaborative E-learning, it is a system described in [40]. It is intended in a blended learning to be used to support teaching and learning processes and also to encourage students to collaborate with peer to learn. The analysis and design phase of this system is done using Beliefs, Desires, Intentions-Agent Based Software Development (BDI-ASPD) to solve a particular problem in agent programming.

This system consists of three main intelligent agents:

- **Student agent**: helps the learning process of students. It provides the mechanism for initialized and update the student's profile and preferences by tracking student behavior. During the learning process, once a student registers to follow a new course, a student agent dedicated to that course is created.
- **Instructor agent**: has the ability to interact with the students as well as allowing a group of students to work on the same assignment. It provides pedagogical materials when requested by Assistant Agent for distributing to students' agents, evaluates the progress and participation of different students, and maintains course progress.
- **Assistant agent**: plays the most essential role in the proposed system. It is initialized as soon as any of the users start to use the system and acts as a mediator between Instructor Agents and Student Agent of a specific course.

It can also track the user's preferences in different areas then can nominee a peer for the user to get help.

MASCE allows the learner to review course materials, ask for help and evaluate the help provided to enable the system to have a list of best assistants [40]. Also, it allows students interacting with their tutors or the other learners using collaboration services provided: forums, wikis, blogs, chat rooms, e-mails. Hierarchical clustering algorithms are used for matching to find the best candidate helper for a peer according to the parameters collected by the system either from the user himself through a questionnaire or through user interaction with the system.

4.1.5. MAS-PLANG

MAS-PLANG is system that aims to offer characteristic of adaptability based on styles of learning for supporting distance adaptive education via the Web [41]. This system is based on the FIPA compliant multi-agent system, using Java, JavaScript Flash and XML languages. It provides content, navigation strategy, and navigation tools according to the students learning style. The environment is composed of two levels of agents: those of the higher level agent programmable (SONIA, Synthetic SMIT, Monitors, and Surfing), and the lower level (didactic agent, user).

- **Sonia Agent:** is a simple reflective agent that for its operation uses data for the tasks that the student or the professor wants it to carry out, as well as certain events happening in the learning environment.
- **Navigation agent:** organizes the navigation paths by direct interaction with databases and by communicating with the didactic agent and the user agent.
- **Synthetic agent:** presents the messages coming from other agents, in the form of suggestions or warnings when the student exhibits special behavior during the learning activity.
- **Supervisor Agent:** monitors the performance of JADE platform and other agents.
- **User Agent:** builds and maintains the student model.
- **Didactic Agent:** is based on information provided by the user agent to select the appropriate pedagogical strategies for the organization of the learning resources.
- **Exercise Action Monitor:** monitors the student during the exercise solving processes.
- **General Action Monitor:** monitors the action of its user and update the user agent.
- **Pedagogic Agent:** evaluates the pedagogic decision rules, whenever student interaction with materials.
- **Controller Agent:** Controls the operation of the agents that are assigned to the students or teacher during the learning session.
- **Exercise Adapter Agent:** Builds the exercise according to the requirement of the student.

4.1.6. SACA

SACA is a Collaborative Learning System based on Agents, it is developed by Lafifi [42]. Its role is to measure the degree of assimilation of knowledge by learners. The different pedagogical activities, including learning, evaluation and of course collaboration are all realized using artificial agents. The assistant agent of the learner can assist him in his learning task (history, statistics, etc.). For its part, the assessment officer takes charge of the process of assessing the degree of assimilation of the knowledge of his learner. In SACA, learner modeling aims to construct a student model based on the observation of its behavior vis-à-vis the interface of a computer system.

SACA is composed of a number of artificial agents:

- **Assistant Learner Agent**: provides the learner with an interface that facilitates the learning task. He holds the learner model of the student, his learning history and his collaboration and evaluation history. The student model provides the information needed to understand learner progress and increase his learning and collaboration opportunities.
- **Pedagogical Agent**: His role is to choose the pedagogical objectives to present to the learner by relying on the state of knowledge of the learner (student model), the final level to be reached (final profile) and the favorable educational strategy.
- **Collaboration Agent or Mediator**: This agent can support the collaborative process between learners (collaboration request, search for a collaborator, choice of collaboration tool, etc.). It can negotiate a possible collaboration with the collaborating agents (mediators) of the other learners.
- **Evaluation Agent**: aims to measure the level of knowledge of the learner by offering a set of exercises of different models and varying difficulties. He is asked to verify the acquisition, by the learner, of the knowledge of an educational objective of the subject to be taught.

4.2. Measurement criteria required for a system

4.2.1. Intelligent Agents Interaction in System

The emergence and development of digital communication tools, which use the Internet network, make possible a faster and more frequent interaction between the tutor and the learner or between the different learners. This interaction is essential to enhance learning communities that enable learners to develop interpersonal skills, and to investigate tacit knowledge shared by community members as well as a formal curriculum. It can deal with the possibility of interaction between humans and agents, checking the possibility of asking for help and collaborating with an appropriate group with a large number of varieties and format that include asynchronous and synchronous communication using text, audio, and video.

4.2.2. Adaptability Measurement of e-Learning Systems

According to the third section, in order to have an effective

educational system, it must no longer be stable with fixed content for all users, but it should respect all the changes made to the user model for building dynamic content. For this reason Brusilovsky [24] has defined the high-level methods for supporting adaptive hypermedia and the lower level techniques that are used to realize or implement this support [1]. Fig.3 and Fig.4 indicate the breakdown of adaptive navigation and presentation techniques appropriate to an adaptive hypermedia system.

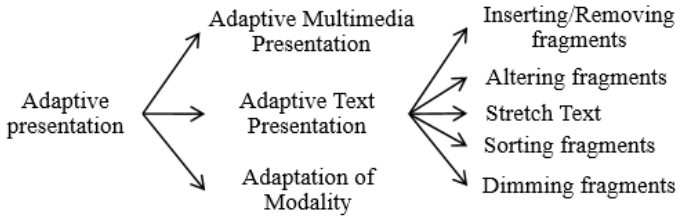


Figure 3 : Taxonomy of adaptive presentation technologies [33].

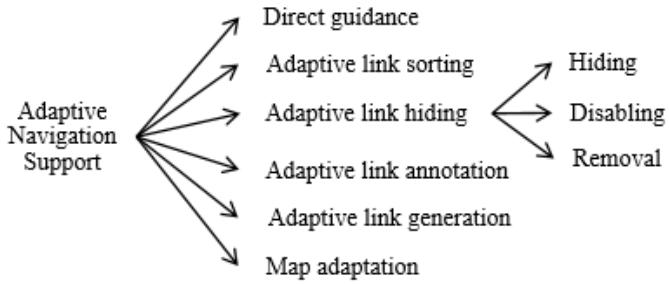


Figure 4: Taxonomy of adaptive Navigation Support technologies [33].

We will use the previously declared systems to analyze their adaptability based on the set of usable techniques.

4.2.3. Cognitive Student Model

This part presents how a student can manage his learning based on the assumption that no one learns exactly the same way. It aims to verify the degree of effectiveness of learner models of existing collaborative intelligent systems based on the nature of the learner characteristics represented. I.e. to check whether it is cognitive characteristics, meta-cognitive or only the representation of demographic information and some tests on the prerequisites of the learner. Cognitive Student Model designates the individual's way of perceiving, evoking, memorizing, and thus understanding the information perceived. This technique aims to use the characteristics defined previously in the learner profile (the duration of reading a page, number of consultation of a course page, the number of repetition of a QCM, etc.) in order to adapt the platform to the learner, to process and store new knowledge.

4.2.4. Learner /group of learners follow up

Unlike face-to-face learning spaces, where the teacher corrects at every moment all information circulating in the classroom during the collective exchange of knowledge. In distance learning environments, it is totally different, sometimes a student may propose involuntarily a false knowledge, or provide knowledge that does not meet the needs of his colleague. Because of a large number of learners connecting to the platform coupled with the high amount of tasks assigned to tutors, sometimes it will be very difficult to check all information shared between learners in different collaborative tools. For those reasons, it is so essential for each learner agent or group of learners' agent to evaluate the help offered by the system or by other learners during the collaboration.

4.2.5. Security Measurement of e-Learning System

Computer security, in general, plays the most important role which is to ensure that an organization's hardware or software resources of an e-Learning system are only used within the intended framework. The aim of security is to protect the most critical information for the conduct of learning activities in order to maintain the trust of learners and also other actors. It is therefore essential to protect them against intrusions and unauthorized access like:

- Protection of passwords, registration procedures and password recovery;
- Access control of learners;
- Content protection against copying;
- Securing exams online;
- Protection of private data related to users;
- Traceability of content, educational activities and sensitive administrative operations through electronic signature and audit mechanisms for highly regulated sectors;
- Backup and restoration of data in the event of a technical incident, etc.

In general, IT security is based on six main objectives:

- **Authentication:** ensuring that only authorized persons have access to resources;
- **Authorization**: provides permission to perform a security function or activity. This security service is often supported by a cryptographic service.
- **Integrity**: aims to ensure that data cannot be modified in an unauthorized or undetected manner;
- **Confidentiality:** ensuring that only authorized persons have access to the resources exchanged. Any unwanted access must be prevented;
- **Non-repudiation:** ensuring that a transaction cannot be denied;
- **Availability:** aims to maintain the proper functioning of the information system and ensure access to installed services and resources with the expected response time;

4.3. Comparison of Existing systems

In the following table, we will analyze and compare some existing environments dedicated to collaborative e-Learning based on multi-agent systems according to the criteria described above.

Table 1: Comparative measurements between existing intelligent collaborative systems

Measurements / Systems		ALLEGRO	BAGHERA	I-MIND	MASCE	MAS-PLANG	SACA
Collaboration Tools	Synchronous	Chat and video Conference	Chat	Chat rooms and Whiteboards	Chat	Chat	Chat
	Asynchronous	Email	X	Forum	Email	Email	Email and Forum
Intelligent actors Interaction		A student agent is interacting with the Teacher tutor agent, Collaborative agent and Expert Agent	A student interacts with three artificial agents: Companion, Tutor and Mediator; A teacher interacts with two artificial agents: Companion and Assistant	The student agent manages the communication channels among students and between the teacher and the students	Each student is interacting with the corresponding Student Manager Agent that helps the learning process of the student	Student interact with different agents (SONIA, Monitoring agents, programmed agents…)	Learner interacts with the Learner Assistant Agent Teacher is interacting with a Teacher Assistant Agent and Teacher's Mediator Agent Tutor is associated with a Tutor Agent
Adaptability Measurement	Navigation	X	Adaptive link annotation	Guidance direct	X	Adaptive link hiding	Adaptive link annotation
	Presentation	Inserting / Removing fragments	X	X	Inserting / Removing fragments	X	X
Cognitive Student Model		Student model contemplates the learning style, understanding of the subjects, limitations and knowledge level of the apprentice.	X	X	MASCE makes it possible to present relatively some preferences such as cognitive style (maximum numbers of concurrent discussions)	X	In this system, the cognitive level represents only the knowledge, and moreover, it is the learner who must manually choose his value (excellent, good, average, etc.).
Learner /group of learners follow up		X	X	I-MINDS student agent evaluates and forms a peer group ("buddy group") for the student that it serves	After each help session, each of the helper and helpee are provided with an evaluation form to evaluate his peer.	X	X
Security Measurement	Authentication	X	The application home page requires users (student, teacher, or administrator) to login to access their own interfaces	Student connect for each of the classrooms using login and password	Student logs to the system and is given a user name and a password for authentication in a specific course.	A student must be connected to the system to have access to revise some specific sections of the lesson, solve a particular problem or enter the chat room.	Learners, authors and tutors must register before any effective use of the system
	Authorization	X	X	X	X	X	X
	Integrity	X	X	X	X	X	X
	Confidentiality	X	X	X	X	X	X
	Non-repudiation	X	X	X	X	X	X
	Availability	X	X	X	X	X	X

4.4. Analysis and prospect for improvements

Reviewing the existing intelligent collaborative systems, we have seen that they have advantages as they have limits. But generally, we can see that they are hardly trying to solve the problem of students' drop out through the integration of certain information and communication techniques to benefit from his strength. If we analyze the table above, we can mention that most of these systems are reinforced by the involvement of many collaborative tools, whether synchronous tools such as chat, or asynchronous tools like email, Whiteboard and forum discussion. Particularly, this can be retained as a great benefit in online learning systems, but this is not always the case. Indeed, the majority of these systems offer certain collaborative tools for learners to work in groups without verification whether this collaboration is beneficial and may meet the needs and preferences of the learner, thus, simply put two people in contact does not necessarily guarantee a good result. Therefore, collaborative tools can have adverse effects if the system does not control the learner behavior. In most cases, learners spend all their time communicating with other colleagues off the topic of learning. For this reason, the system must be managed against the misuse of collaboration tools so that it does not lose its educational value. Further, the majority of these systems are interested in collaborative learning in groups, but they have never analyzed how these groups should be constituted even there have been several studies conducted on the ideal size of the group, and methods of choosing members for regrouping.

As we have already shown in the third section of this article, the integration of agents in an e-Learning system suggests a very important improvement. It makes agents able to diminish and facilitate tasks to all human actors of learning platform. In order to have a more efficient learning environment, it is recommended to separate the tasks of different actors in the platform. I.e. it is better to have a system with multiple agents where everyone is responsible for specific tasks, instead of having, for example, one tutor or teacher agent who take overall responsibility for managing the whole system. So, in an e-Learning system, it's better to separate the teacher role for two or more actors such as the pedagogical author and the expert author. Where the pedagogical author (or instructional designer) utilizes all the pedagogical design principles, models, and learning theories available to achieve pedagogical goals as he supports the management and development of e-Learning course content, the creation of storyboards, development compliance manuals, etc. While the expert author tries to define what needs to be added to the e-Learning course, and what can be put aside. Most of the time, a different expert author is used for each new e-Learning course. He usually works closely with the instructional designer to determine the key points and learning objectives of the course, as well as how the content will be delivered. Similarly to the tutor agent, it is necessary to separate the maximum tasks by creating specific agents such as the filtering agent that allows filtering multimedia documents stored in e-Learning system's database in order to select an appropriate educational content adapted to the criteria specified in the student model. Or the technical assistant agent that is responsible for the good technical functioning, and the necessary updates in the learning environment. He must always be available to help learners if they have a technical problem, or if they do not master the use of the computer tools.

From the table 1, we can observe that the personalization applied in these systems does not exceed more than one aspect of adaptation, i.e. they only adapt the navigation or the presentation of the interfaces. Moreover, they are limited to the use of just one adaptation technique, which is often inserting and removing fragment for presentation adaptation, and the adaptive annotation for navigation adaptation Indeed, this adaptation remains insufficient in the most existing systems and make their interfaces so very poor ergonomically. So we can summarize that there is a lack in the adaptation that should not be left behind.

This analyze, we noted that the problem of the lack of adaptation techniques used is not the only obstacle that interferes with the flexibility of adaptation of those systems, but there is also the lack of learner modeling. Most distance learning systems have been structured with a minimum of consideration of relevant aspects of human cognitive architecture [43]. Each of them does not perform any user model representations while others offer a limited learner model such as SACA for example. It deals only with the acquisition of learner's knowledge whereas learner modeling is no longer limited to modeling knowledge but also the representation of psychological aspects such as information on the emotional state and emotions, the intentions of the learner. In this system, the cognitive aspects can only be dealt with a simple question asked to the user aiming to specify the level of mastery of concept taught. Before starting the learning activity, the learner must manually indicate his level of knowledge using a simple value (excellent, good, average ...), which is still irrelevant. For the above reasons, it's important to think about improving existing systems or to design a new system based on cognitive modeling of the learner, i.e. it aims to represent as closely as possible all the cognitive and metacognitive aspects relating to the behaviors of a learner.

Not only misuse of cognitive student model but also the problem of helps evaluation by Learner agent or group of learner agent. We can distinguish that MASCE allows a learner to ask for help and to evaluate the help provided by another learner, and I-MINDS allows the teacher to monitor the activities of learners and groups. And both systems are used to evaluate learners. Unlike MASCE, I-MINDS offers users the possibility of forming a group [39]. However, the others do not offer any follow-up of learners.

Concerning the systems analyzed above, we can notice that the most of them lack the most basic conditions to prevent intruders. While the evolution of educational systems via the Internet currently is leading to a new concept of security threats against applications and users that affect e-Learning content design and tools. Besides that, even other systems that are partially secure, they only protect the authentication problem like MASCE, MAS-PLAN, and I-MINDS. Indeed, when designing a secure e-Learning system, designers must have a clear idea of the threats they must prevent, and the technical capabilities of the attackers to specify accordingly the preliminary steps required to secure the system [44]. They must take into consideration the identification verification of the student being assessed by ensuring that the person answering is the learner who must have passed the assessment and not another person. They must also protect the e-

Learning systems, especially when passing evaluation, against nodes and links attacks which affects the availability of evaluation pages and other resources [44]. In the same way, the system must be protected by preventing unwanted access and theft of documents that reveal user privacy information. In addition, they must secure the evaluation procedures by ensuring that the learners did not have access to the assessment procedures before taking the exam. Indeed, in an e-Learning environment, depending on time zones, some students can take their exam in one country while others are still reviewing. This problem does not matter for environments that just have awareness training on a subject, but they are more important for certified training. And as a result, that will affect the reputation of the risky platform, and negatively affect the quality of the degrees acquired. For that, the most likely attacks for these systems should be described independently of the specific implementation, to help designers to eliminate or mitigate these attacks if possible in the design phase not waiting for actual attacks to occur.

5. Conclusion

According to the originality of solutions offered by adaptive multi-agent systems, recently, e-Learning researchers often use MAS to solve many problems as selecting autonomously the most suitable learning objects that meet students' preferences, or collecting and processing learners' data in order to monitor their progress, motivate and guide them, and avoid their abandonment. In this study, we present the interest of artificial intelligence in e-Learning system for the purpose of controlling the students' dropout phenomenon. We analyzed some existing intelligent collaborative e-Learning based on various characters in order to determine their limits and propose some improvement that can make them more reliable for use.

In this paper which represent an analysis of some existing intelligent collaborative systems, we concluded that these systems are still in need of improvement at the level of integration of a powerful learner model. This model should not be limited to a simple collection of learner biographical information or an assessment of learner level knowledge, but it must also be able to represent psychological aspects such as information on affective state and emotions, learner intentions, as well as cognitive and metacognitive characteristics. We found also that the most of the analyzed systems do not offer any follow-up of learners. Moreover, the lack of the most basic conditions to prevent intruders even the other systems that are partially secure, they only protect the authentication problems.

References

[1]. CLARK, Ruth C. and MAYER, Richard E. E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. John Wiley & Sons, 2016.

[2]. OYE, N. A., IAHAD, N., MADAR, M. J., and al. The impact of e-learning on students' performance in tertiary institutions. International Journal of Computer Networks and Wireless Communications, 2012, vol. 2, no 2, p. 121-130.

[3]. BELANGER, Yvonne and THORNTON, Jessica. Bioelectricity: A quantitative approach Duke University's first MOOC. 2013.

[4]. HALAWA, Sherif, GREENE, Daniel, and MITCHELL, John. Dropout prediction in MOOCs using learner activity features. Experiences and best practices in and around MOOCs, 2014, vol. 7, p. 3-12.

[5]. ONAH, Daniel FO, SINCLAIR, Jane, and BOYATT, Russell. Dropout rates of massive open online courses: behavioural patterns. EDULEARN14 proceedings, 2014, p. 5825-5834.

[6]. GÜTL, Christian, RIZZARDINI, Rocael Hernández, CHANG, Vanessa, and al. Attrition in MOOC: Lessons learned from drop-out students. In: International Workshop on Learning Technology for Education in Cloud. Springer, Cham, 2014. p. 37-48. https://doi.org/10.1007/978-3-319-10671-7_4

[7]. GEE, Sue. MITx-the fallout rate. Message posted, 2012, vol. 16.

[8]. BARKLEY, Elizabeth F., CROSS, K. Patricia, and MAJOR, ClaireH. Collaborative learning techniques: A handbook for college faculty. John Wiley & Sons, 2014.

[9]. ASSELMAN, Amal, NASSEH, Az-Eddine, and AAMMOU, Souhaib. Survey of intelligent collaborative E-learning systems. In: Emerging eLearning Technologies and Applications (ICETA), 2017 15th International Conference on. IEEE, 2017. p. 1-7. 10.1109/ICETA.2017.8102463

[10]. ZAPPAROLLI, Luciana Silva and STIUBIENER, Itana. FAG-a management support tool with BI techniques to assist teachers in the virtual learning environment Moodle. Advances in Science, Technology and Engineering Systems Journal, 2017, vol. 2, no 3, p. 587-597. 10.25046/aj020375

[11]. DE KOCH, Nora Parcus. Software engineering for adaptive hypermedia systems. Ph. DThesis, Verlag Uni-Druck, Munich, 2001.

[12]. JÄRVELÄ, Sanna, KIRSCHNER, Paul A., PANADERO, Ernesto, and al. Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. Educational Technology Research and Development, 2015, vol. 63, no 1, p. 125-142. https://doi.org/10.1007/s11423-014-9358-1

[13]. ROSCHELLE, Jeremy. Learning by collaborating: Convergent conceptual change. The journal of the learning sciences, 1992, vol. 2, no 3, p. 235-276. https://doi.org/10.1207/s15327809jls0203_1

[14]. WALCKIERS, Marc and DE PRAETERE, Thomas. L'apprentissage collaboratif en ligne, huit avantages qui en font un must. Distances and savoirs, 2004, vol. 2, no 1, p. 53-75. 10.3166/ds.2.53-75

[15]. MCLAREN, Bruce M., SCHEUER, Oliver, and MIKŠÁTKO, Jan. Supporting collaborative learning and e-Discussions using artificial intelligence techniques. International Journal of Artificial Intelligence in Education, 2010, vol. 20, no 1, p. 1-46.

[16]. MCCARTHY, John, MINSKY, Marvin L., ROCHESTER, Nathaniel, and al. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 2006, vol. 27, no 4, p. 12.

[17]. RUSSELL, Stuart J. and NORVIG, Peter. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,, 2016.

[18]. WENGER, Etienne. Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge. Morgan Kaufmann, 2014.

[19]. BRUSILOVSKY, Peter. Methods and techniques of adaptive hypermedia. User modeling and user-adapted interaction, 1996, vol. 6, no 2-3, p. 87-129. https://doi.org/10.1007/BF00143964

[20]. MIHALCA, Rodica, UTA, A., ANDREESCU, Anca, and al. Knowledge management in E-learning systems. Revista Informatica Economică, 2008, vol. 2, no 46, p. 60-65.

[21]. GHAHRAMANI, Zoubin. Probabilistic machine learning and artificial intelligence. Nature, 2015, vol. 521, no 7553, p. 452.

[22]. WOOLDRIDGE, Michael. An introduction to multiagent systems. John Wiley & Sons, 2009.

[23]. RUSSELL, Stuart and NORVIG, Peter. Intelligence artificielle: Avec plus de 500 exercices. Pearson Education France, 2010.

[24]. CHAIB-DRAA, Brahim, JARRAS, Imed, and MOULIN, Bernard. Systèmes multi-agents: principes généraux and applications. Principes and architectures des systèmes multi-agents, 2001.

[25]. LABIDI, Sofiane and LEJOUAD, Wided. De l'intelligence artificielle distribuée aux systèmes multi-agents. 1993. Thèse de doctorat. INRIA.

[26]. FERBER, Jacques. Les systèmes multi-agents: vers une intelligence collective. InterEditions, 1995.

[27]. ASSELMAN, Amal, AAMMOU, Souhaib, and NASSEH, Az-Eddine. Comparative study of cognitive architectures. International Research Journal of Computer Science (IRJCS), 2015, vol. 2, no 9.

[28]. ANDERSON, John R. How can the human mind occur in the physical universe?. Oxford University Press, 2009.

[29]. KOTSERUBA, Iuliia, GONZALEZ, Oscar J. Avella, and TSOTSOS, John K. A review of 40 years of cognitive architecture research: focus on perception, attention, learning and applications. arXiv preprint arXiv:1610.08602, 2016.

[30]. BISWAS, Pradipta and SPRINGETT, Mark. User Modeling. The Wiley Handbook of Human Computer Interaction, 2018, vol. 1, p. 143-169. https://doi.org/10.1002/9781118976005.ch8

[31]. VAN RIJN, Hedderik, JOHNSON, Addie, and TAATGEN, Niels. Cognitive user modeling. Handbook of human factors in web design, 2nd edn. CRC Press, New Yersey, 2011, p. 527-542. http://doi.org/doi:10.1201/b10855-34

[32]. BRUSILOVSKY, Peter and MILLÁN, Eva. User models for adaptive hypermedia and adaptive educational systems. In: The adaptive web. Springer

Berlin Heidelberg, 2007, vol. 4321, p. 3-53. https://doi.org/10.1007/978-3-540-72079-9_1
[33]. BRUSILOVSKY, Peter. User modeling and user-adapted interaction. 2001.
[34]. CHEN, Sherry Y. and MAGOULAS, George D. (ed.). Adaptable and adaptive hypermedia systems. IGI Global, 2005.
[35]. AL-AZAWEI, Ahmed and BADII, Atta. State of the art of learning styles-based adaptive educational hypermedia systems (LS-BAEHSs). International Journal of Computer Science & Information Technology, 2014, vol. 6, no 3, p. 1. http://dx.doi.org/10.5121/ijcsit.2014.6301
[36]. KORB, Kevin B. and NICHOLSON, Ann E. Bayesian artificial intelligence. CRC press, 2010.
[37]. VICCARI, Rosa M., OVALLE, Demetrio A., and JIMÉNEZ, Jovani A. ALLEGRO: Teaching/learning multi-agent environment using instructional planning and cases-based reasoning (CBR). CLEI Electronic Journal, 2007, vol. 10, no 1, p. 1-20.
[38]. WEBBER, Carine, BERGIA, Loris, PESTY, Sylvie, and al. The Baghera project: a multi-agent architecture for human learning. In: Workshop-Multi-Agent Architectures for Distributed Learning Environments. In Proceedings International Conference on AI and Education. San Antonio, Texas. 2001.
[39]. SOH, Leen-Kiat, KHANDAKER, Nobel, and JIANG, Hong. IMINDS: a multiagent system for intelligent computer-supported collaborative learning and classroom management. International Journal of Artificial Intelligence in Education, 2008, vol. 18, no 2, p. 119-151.
[40]. MAHDI, Hani and ATTIA, Sally S. MASCE: a multi-agent system for collaborative e-learning. In: Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on. IEEE, 2008. p. 925-926. 10.1109/AICCSA.2008.4493647
[41]. PEÑA, Clara-Inés, MARZO, Jose-L., and DE LA ROSA, JosepLluis. Intelligent agents in a teaching and learning environment on the Web. In: Proceedings of the international conference on advanced learning technologies. Palmerston North, NZ: IEEE Learning Technology Task Force. 2002. p. 21-27.
[42]. LAFIFI, Yacine. SACA: un Système dApprentissage Collaboratif. 2007. Thèse de doctorat. PhD Thesis, Computer science department, University of Annaba, Algeria.
[43]. BALAŽ, Zdenko and PREDAVEC, Davor. Cognitive Cybernetics vs. Captology. Advances in Science, Technology and Engineering Systems Journal, 2017, vol. 2, no 6, p. 107. 10.25046/aj020614
[44]. NICKOLOVA, Maria and NICKOLOV, Eugene. Threat model for user security in e-learning systems. International Journal Information Technologies and Knowledge, 2007, vol. 1, no 1, p. 341-347.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 67-73 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Similarity-based Resource Selection for Scientific Workflows in Cloud Computing

Takahiro Koita[1,*], Yu Manabe[2]

[1]*Doshisha University, Faculty of Science and Engineering, Japan*

[2]*NAIST, Department, Division of Information Science, Japan*

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 22 October, 2018
Online: 01 November, 2018

Keywords:
Cloud computing
Amazon web services
Scientific workflows
Resource selection

ABSTRACT

There are high expectations for commercial cloud services as an economical computation resource when executing scientific computing workflows, for which the computation is increasing on a daily basis. However, no method has been developed for determining whether a scientific computing workflow can be executed at a low usage cost, and thus scientists have difficulty in selecting from the diverse range of computational resources. The aim of this study is to provide clear criteria for selecting a computational resource while executing a scientific computing workflow. This study focuses on the performance of application execution for one such commercial cloud service, Amazon EC2, and proposes a method for selecting the optimal resource showing high similarity to a target application in execution time and usage cost. The novelty of this study is its approach of employing application similarity in resource selection, which enables us to apply our method to unknown applications. The contributions of this work include (1) formularizing performance values of computational resources, as well as similarity values of applications, and (2) demonstrating the effectiveness of using these values for resource selection.

1. Introduction

This paper is an extension of work originally presented in ICBDA2018 [1]. Scientific computing workflows [2] are applications that performs a sequence of processes by dividing applications handling scientific computing into small tasks and, by executing these tasks in stages. Figure 1 shows an overview of scientific computing workflows of Epigenomics and Motage. Characteristically, scientific computing workflows can deal with large amount of data, and the work quantity differs for each task. Here, as examples, we introduce three types of scientific computing workflows. Montage is an application developed by NASA that processes celestial images. A feature of Motage is that, since it handles large-sized images, it requires high levels of I/O performance. Broadband is an application that generates a vibration record diagram from multiple earthquake simulations, and requires high levels of memory performance. Epigenomics is an application dealing with DNA, and requires high levels of processing ability based on CPU performance. Thus far, scientific research has mainly consisted of experiments and theories. However, with developments in hardware, advanced calculation

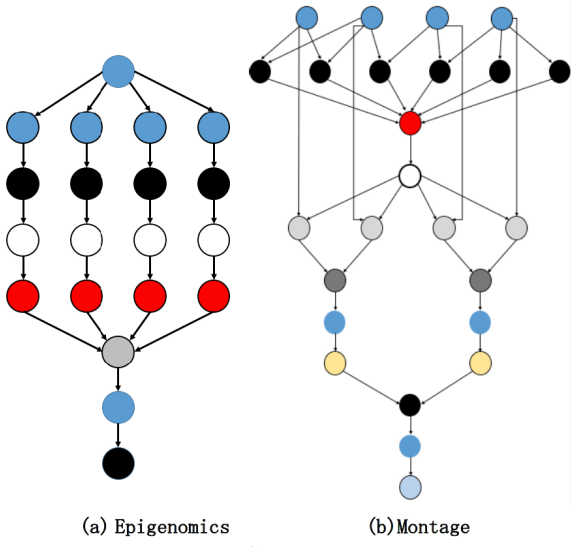


Figure 1. Scientific Workflow

*Takahiro KOITA, Email: tkoita@gmail.com

Table 1. List of Instances

(a) Role list

role	instance name	
General purpose	t2	m4
CPU optimized	c4	
Memory optimized	r3	x1
Storage optimized	i2	d2
GPU instance	g2	
GPU computing	p2	

(b) Performance list

performance			
nano	micro	small	medium
large	xlarge	2xlarge	4xlarge
10xlarge	16xlarge	32xlarge	

Table 2. Instance Performance

	vCPU	ECU	memory (GB)	storage (GB)	cost ($/h)
t2.micro	1	variable	1	8	0.013
c4.large	2	8	3.75	8	0.105
m4.large	2	6.5	8	8	0.12
m4.xlarge	4	13	16	8	0.239
r3.large	2	6.5	15.25	32	0.166
i2.xlarge	4	14	30.5	800	0.853

has become possible, and computer-based simulations have become an essential new research method. The importance of scientific computing workflows are only expected to grow for future science. Many scientific computing workflows have been developed based on distributed processing using high performance computers (HPC) with grids, PC clusters, and supercomputers, and scientists have used their own PC clusters and grid computing, such as Open Science Grid [3], when executing scientific computing workflows. However, through the development of hardware, the data quantity that can be operated is increasing on a daily basis, and the processing capabilities of computation sources and the storage capacity required are expanding in the same way.

When executing scientific computing workflows, for which the computational data is increasing on a daily basis, the use of commercial sloud services as a computation resource, in place of PC clusters and Open Science Grid, has attracted attention. The commercial cloud service is a service in which scientists can use servers on the network by paying usage fees. Features of such services are that computational sources and storage can be swiftly added, and computational sources of various performance (instances) are prepared. Scientific computing workflows are designed based on distributed processing, and it is possible to execute these in the cloud in which distributed processing is performed through distributed computing. Additionally, as a wide variety of performance instances are prepared, tasks with different processing can be performed in respectively optimized environments. It also has the characteristics of being a measured rate system in which you only pay for the time you use, the fact that maintenance costs are not required, and initial investment for constructing facilities is not necessary. It promises to be applicable to scientific computing workflows, and is to be used as a highly-economical computation source. In this study, we use the commercial cloud service Amazon Elastic Compute Cloud (EC2) [4] used in the preceding research [5]. EC2 is a web service provided by Amazon. The users can select virtual machines, called instances, according to various purposes. With EC2, five types of roles and multiple respective processing resources are prepared. A list of the instances is shown in Table 1.

One of the important problems with using EC2 is that it is difficult to select the instance and the application to execute with the instance performance table. If the instance performance does not satisfy the performance requirements of the application, execution will be impossible, or the execution time will increase, leading to an increase in usage costs. On the other hand, if the instance performance is higher than necessary, the cost per unit time will be higher, and even if the execution time is shorter, the costs would increase. Currently, when selecting the instance to execute the application, specialized knowledge about applications and cloud or user experience are required. This stiuation makes it difficult to select a suitable computational resource from a large number of computation resources when considering execution time and usage costs, and this is the problem for scientists using commercial cloud services.

To solve the problem, this study aims to provide a clear criterion for selecting instances for executing scientific computing workflows. Using the provided selection criteria, it will be possible for scientists to casually engage in cloud services, and to perform experiments using advanced computing resources for low research fees. The novelty of this study is its approach of employing application similarity in resource selection, which enables us to apply our method to unknown applications. The contributions of this work include (1) formularizing performance values of computational resources, as well as similarity values of applications, and (2) demonstrating the effectiveness of using these values for resource selection.

2. Current Issues

There are four main issues in executing scientific computation workflow using the cloud, as follows:

1) Virtualization overhead

2) Low throughput in shared/parallel file systems

3) Low network performance

4) Unclear usage costs

Issue 1 appears in a significant way when CPU performance is required. Additionally, 2 and 3 clearly have an impact on applications requiring I/O performance [6]. Based on the above features, the current commercial cloud services cannot achieve HPC-equivalent performance. It is expected that issues 1 to 3 shall be resolved on the hardware front, through the development of commercial cloud services. However, this will not solve 4, which occurs when scientists use commercial cloud services. Scientists need to select the computation resources satisfying the processing ability required for the application based on uncertain factors such as their own knowledge and experience. A cause of issue 4 not being satisfied is that there are no criteria for selecting computational resources that consider the necessity of applications [7].

This study focuses on the issue of usage cost. The issue is how to select instances that will satisfy computing performance requirements and have the lowest user cost when executing a scientific workflow on a commercial cloud service. In EC2, the performance of each instance is published, and users can determine the computation resources based on the performance table values. Table 2 shows part of the published performance table. vCPU expresses the number of virtual Server cores, and ECU (EC2 Computing Unit) is a numerical representation of the total processing performance for the instants. In case of an instance where ECU is 8 and vCPU is 2, the CPU processing ability per core is 4. EBS (Elastic Block Storage) is the block unit storage provided by Amazon, and a total of four types are prepared, comprising two types of SSD and two types of HDD [8]. For all instances in this study, the versatile SSD type found in the default settings is used.

Currently, the user selects the instance using their experience, based on the performance capability required for the executed applications and the values of the instance performance table, and it is possible that the instance with the shortest execution time or the lowest usage costs may not be selected.

Table 3. Instance Performance Values

	ECU	EMU	EFU
m4.large	6.50	9.30	9.50
m4.xlarge	13.0	9.62	9.25
c4.large	8.00	8.00	8.00
c4.xlarge	16.0	8.04	9.08
r3.large	6.50	8.48	9.05
r3.xlarge	13.0	8.45	9.05

We explain this situation using the example of a prime number calculation application. Prime number calculation applications are applications that mainly require CPU processing ability. For this reason, it is predicted that the user will select the c4 interface, which has enhanced CPU performance. However, at that time, they need to decide whether to choose the c4.large with 8 ECU, or the c4.xlarge with 16 ECU. The result of actually executing this was that the execution time was shorter for c4.xlarge, but the usage costs were lower with c4.large. Due to this, until we actually execute the application, it is unclear which instance has the shortest execution time or which has a lower usage costs. Additionally, the processing ability used for prime number calculation examples is virtually CPU only, but with the actual application, memory and I/O processing ability are required at the same time. When selecting the instance, it is important to have a proper understanding of the processing ability required by the application.

To execute the application with a short execution time or a low usage cost, it is necessary to quantitatively grasp the performance ability required by the application and the instance performance and clarify these relationships. Therefore, in the next section, we will quantitatively demonstrate the instance performance and the processing ability required by the application and perform preliminary experiments to provide clear selection criteria.

Very few studies have been made to quantitatively grasp the performance required by the application or the instance. Tovar et al. [9] classified tasks in scientific workflows and proposed an estimation method for the tasks. They showed that the execution time can be estimated and that CPU, memory and I/O performance indexes are important for this estimation. Sfiligoi et al. [10] showed the characteristics of scientific workflows statistically, and the results were effective for their experiment's applications. However, these studies are useful only for known applications whose behavior information can be given well in advance of instance selection. Consequently, if such information is insufficient, these studies cannot be applied. Thus, the current study employs several values to achieve resource selection for unknown applications. Furthermore, previous studies assumed that their target instance was a single type and thus did not consider the various types of instances in commercial cloud services.

3. Preliminary Experiment

We perform preliminary experiments to quantitatively show the instance performance and processing ability required by the application.

3.1. Instance Performance Value

We describe the instance performance as numerical values. We focus on instance performance in terms of CPU, memory, and I/O. This is because CPU, memory, and storage are enhanced respectively in EC2, and because the instances are mainly prepared in relation to these, it is assumed that these will have the greatest impact on execution time and usage cost. The CPU processing ability uses ECU, published by Amazon. In this study, memory and I/O processing ability are defined respectively as EMU and EFU, and these are measured and expressed numerically in these preliminary experiments. In these preliminary experiments, the versatile instances m4.large and m4.xlarge，the CPU optimization instances c4.large and c4.xlarge，and the memory optimization instances r3.large and r3.xlarge are used. The performance of each of these is shown, respectively, in Table 3.

Table 4. Execution Performance Values
(partial result only for two applications in [11])

	execution time [sec]	CPU	memory	I/O
Memory bound apl.	291.6	44.9	31.4	30.7
I/O bound apl.	25.6	3.94	2.75	2.70

Measurements are performed using a program prepared for the purpose of measuring performance evaluations. Memory performance evaluations involve reading and writing memory multiple times, whereas I/O performance is assessed by reading and writing a text file multiple times. The respective execution times are measured, with the ratio with the c4.large value and 8, which is the same as ECU, to discover the EMU and EFU of each instance. The respective instance performance is shown in Table 4.

3.2. Execution Performance Value

In this experiment, the processing ability required by the application expressed as a numerical value is used as the execution performance value. The execution performance value shows the effect of the CPU, memory, and IO required for the application on the execution time. As an example, we shall show a formula for obtaining the execution performance value based on CPU performance. The performance values based on memory or I/O performance can be calculated by making the respective ECU values the EMU and EFU values according to the following formula.

Here, as an example, we shall seek the two application execution performance values of memory performance evaluation and I/O performance evaluation used when evaluating instance performance. The execution performance values need to actually be executed with the instances. In this preliminary experiment, measurement was performed using the versatile instance m4.large. The execution time and execution performance values based on the CPU, memory, and I/O performance are as follows.

From the results, we can see that the size of the execution performance values changes depending on the execution time, and that differences appear in the execution performance value ratios based on the processing content. The two applications used in this preliminary inspection have high execution performance values based on CPU performance and, as with the instance performance evaluation, the execution time for both applications was shorter with the c4 instance optimized for CPU performance, by referencing the execution performance values, it is possible to grasp the processing capability required by the application.

4. Proposed Method

We propose a method for selecting resources that uses an application with similar execution performance to select the instance that can run an application in the lowest time or with lowest usage cost. In the past, the run time and usage cost were unknown before actually running an application, and there was a risk of costs increasing when the application was run several times. The proposed method enables a resource to be selected, running the application a minimum number of times, by considering the application execution performance values and instance performance.

We expect that if the performance required by two applications is the same, the computing resources required for the shortest execution time or lowest usage cost will be the same for them as well. The proposed method selects an application with similar execution performance values as the application in question from among several that have been run in the past, and selects the computing resource able to execute the application in question in the shortest time or at lowest cost. The method is comprised of the following four steps.

1) Standardization Step: Measure the execution performance of the sample applications

2) Measurement Step: Measure the execution performance of the application in question

3) Comparison Step: Select a sample application with similar execution performance values

4) Selection Step: Select the instance with shortest execution time or with lowest usage cost

Details of each step are described below.

Standardization Step: The sample applications are executed on each instance, and the instances producing the shortest run time and lowest usage cost are selected. The execution performance is also computed using an arbitrarily selected instance. Several applications performing different processes are used as sample applications for the proposed method.

Measurement Step: This step deals with the application for which a computing resource is being selected. The application is executed on an instance selected in the standardization step to measure its execution performance values.

Comparison Step: In this step, the execution performance values of all sample applications measured in the standardization step are compared with the execution performance values of the application in question, as measured in the measurement step. For this method, a similarity level is used for this comparison. The similarity level is expressed as a distance between the execution performance values of the two applications. The normalized execution performance values of CPU, memory and I/O of application A are denoted $E_{Ae}, E_{Am},$ and E_{Ai}, respectively. These execution peformance values, are obitained by the preliminary experiment described in Section 3. Similarly, the execution performance values for application B are denoted $E_{Be}, E_{Bm},$ and E_{Bi}. The similarity, D_{AB}, is given by the following equation (1). This equation uses Euclidean Distance between applications A and B. If the similarity value is sufficiently high, they are considered similar applications. The highest similarity value is thus used to select the most similar application. To select a resource, the target application A is fixed while application B varies. That is, the distances to application A from all other applications are calculated. The distance is determined by the values of execution time, memory, and I/O described in the previous section. More details can be seen in an earlier work [11].

$$D_{AB} = \sqrt{(E_{Ae} - E_{Be})^2 + (E_{Am} - E_{Bm})^2 + (E_{Ai} - E_{Bi})^2} \quad (1)$$

Selection Step: In the selection step, the similarity of the application with each of the sample applications is computed. The sample application with the smallest similarity level is selected as the one that is most similar to the application in question. The instance able to execute this most-similar sample application with the shortest execution time or lowest cost is then selected as the instance to run the application in question.

5. Evaluation

To show that instances can be selected based on similarity of execution performance values, we conducted experiments to evaluate the effect of similarity on the execution time and usage cost.

5.1. Experiment Overview

Using the same instances as in the preliminary experiments, the execution time, usage cost and similarity were measured for multiple applications. If the instances running the applications with the lowest execution time and usage cost are the same for applications that are similar, the proposed method will select resources correctly. The instances used were the same as those used in the preliminary experiments. The test procedure was as follows:

1) Run applications

2) Calculate execution performance values

3) Calculate similarities

4) Select similar applications.

Each of these steps is described in more detail below.

Run applications: Multiple applications were run on each instance, and the execution time was measured. From the execution times for each instance, the instances producing the shortest execution time and lowest usage cost were selected for each.

Calculate execution performance values: Each application was run on an arbitrarily selected instance and the execution performance values were measured. For these tests, we used the m4.large general-purpose instance.

Calculate similarities: Here, the computed execution performance values were normalized, and the similarity to all of the other applications was computed for each application.

Select similar applications: For each application, the one with the lowest similarity level was selected as the most similar application. Each application was compared with the other application most similar to it, and we checked whether the instances producing the shortest execution time and usage cost were the same.

5.2. Applications

The applications used here include Sysbench [12] and UnixBench [13], which are comprehensive benchmark applications, and Hadoop [14], which is a distributed framework. These are described in more detail below.

Sysbench is a general benchmark application for Linux/Unix operating systems. Sysbench has six types of evaluation (e.g. CPU or memory) and enables each of them to execute with adjusting application parameters such as the number of CPUs or the file size. For our experiments, we performed CPU, memory and I/O tests. Prime numbers are computed for the CPU test, reading and writing to memory is done for the memory test, and reading and writing files to storage is done for the I/O test. Each test was done repeatedly and the execution times were measured. The term of test means a specific execution to perform one type of evaluation.

UnixBench is a benchmark application used with Unix-type operating systems. The test covers a variety of tasks from integer arithmetic through to OS system calls. In these experiments, benchmarks for integer computation (Dhrystone), floating-point computation (Whetstone), and file copying (fsdisk) were performed. Results are given in terms of processing capability per unit time. These were converted to results in terms of a time required to complete a fixed-size process for these experiments.

Hadoop is a distributed framework that enables multiple computers to be treated as a single computer with improved performance. Hadoop can be used in any of three modes: stand-alone mode, which runs on a single CPU, pseudo-distributed mode, which virtualizes use of two machines on a single machine, and fully-distributed mode, which uses multiple computers. For these experiments, we used pseudo-distributed mode, measuring execution time for standard sample processes including computing pi, counting words, and sorting files.

5.3. Results

Each application was executed on each of the instances. For some of the applications, such as Dhrystone in UnixBench, the results are given in number of loops per second rather than total execution time. In such cases, the results were converted to a time required to perform a set number of loops. Execution times and usage costs are given in Tables 4 and 5, and execution performance values, as discussed previously, are given in Table 6.

Instance c4.xlarge had the shortest execution time, and instance c4.large had the lowest usage cost in most cases. Execution performance values were calculated using the execution times and the ECU, EMU, and EFU performance values for m4.large, and these were then normalized.

Similarities were then computed using these values. Below, we give an example of computing the similarity, D_{pw}, is given by equation (2) using the execution performance values from the prime number computation in Sysbench and the word count process on Hadoop.

$$D_{pw} = \sqrt{(0.038 - 0.020)^2 + (0.047 - 0.025)^2 + (0.047 - 0.025)^2} \quad (2)$$

Similarities were computed for all process pairs, and that with the smallest similarity value was designated as the similar application for each application. Table 8 shows whether the instances producing the shortest execution time and lowest usage cost were the same for these similar applications.

The process most similar to the prime number computation on Sysbench was word count on Hadoop. The instance with the shortest execution time for the prime number process in Sysbench was c4.xlarge, which was the same as for the Hadoop word count,

Table 5. Execution Times [sec]

application		execution time					
		m4.large	m4.xlarge	c4.large	c4.xlarge	r3.large	r3.xlarge
Sysbench	prime number	157.51	78.99	128.68	64.32	153.31	76.47
	memory read	42.55	38.61	35.79	33.72	40.62	36.39
	memory write	52.32	41.82	43.63	36.96	49.90	40.09
	random read/write	2.07	0.42	1.71	0.78	1.51	1.00
	sequential read/write	34.42	22.90	34.36	22.90	38.14	24.60
hadoop	pi	3328.29	1692.45	2828.35	1438.95	3339.85	1698.05
	word count	83.14	54.62	77.71	50.90	88.76	56.86
	file sort	40.86	36.68	40.00	36.36	42.82	36.55
UnixBench	Dhrystone	0.76	0.85	0.64	0.32	0.74	0.37
	Whetstone	0.34	0.73	0.35	0.17	0.41	0.20
	fsdisk	0.71	0.76	0.68	0.83	0.95	1.14

Table 6. Usage Costs [$]

application		usage cost					
		m4.large	m4.xlarge	c4.large	c4.xlarge	r3.large	r3.xlarge
Sysbench	prime number	21.89	21.96	16.21	16.21	30.66	30.51
	memory read	5.91	10.73	4.51	8.50	8.12	14.52
	memory write	7.27	11.63	5.50	9.31	9.98	15.99
	random read/write	0.29	0.12	0.22	0.20	0.30	0.40
	sequential read/write	4.78	6.37	4.33	5.77	7.63	9.81
Hadoop	pi	462.63	470.50	356.37	362.62	667.97	677.52
	word count	11.56	15.18	9.79	12.83	17.75	22.69
	file sort	5.68	10.20	5.04	9.16	8.56	14.58
UnixBench	Dhrystone	0.11	0.24	0.08	0.08	0.15	0.15
	Whetstone	0.05	0.20	0.04	0.04	0.08	0.08
	fsdisk	0.10	0.21	0.09	0.21	0.19	0.46

Table 7. Execution Performance Value

application		performance value		
		CPU	memory	I/O
Sysbench	prime number	0.03837	0.04725	0.04725
	memory read	0.01030	0.01270	0.01270
	memory write	0.01269	0.01564	0.01564
	random read/write	0.00042	0.00054	0.00054
	sequential read/write	0.00832	0.01026	0.01026
Hadoop	pi	0.81248	1.00000	1.00000
	word count	0.02021	0.02490	0.02490
	file sort	0.00989	0.01220	0.01220
UnixBench	Dhrystone	0.00010	0.00015	0.00015
	Whetstone	0.00000	0.00002	0.00002
	fsdisk	0.00009	0.00013	0.00013

Table 8. Matching Instance and Result of Similarity-based method

application		best time	best cost	similarity application	best time instance	best cost instance
Sysbench	prime number	c4.xlarge	c4.xlarge	word count	Same	Different
	memory read	c4.xlarge	c4.large	file sort	Same	Same
	memory write	c4.xlarge	c4.large	memory read	Same	Same
	random read/write	m4.xlarge	m4.xlarge	Dhrystone	Different	Different
	sequential read/write	m4.xlarge	c4.large	file sort	Different	Same
Hadoop	pi	c4.xlarge	c4.large	prime number	Same	Different
	word count	c4.xlarge	c4.large	memory write	Same	Same
	file sort	c4.xlarge	c4.large	memory read	Same	Same
UnixBench	Dhrystone	c4.xlarge	c4.xlarge	fsdisk	Different	Different
	Whetstone	c4.xlarge	c4.large	fsdisk	Different	Same
	fsdisk	c4.large	c4.large	Dhrystone	Different	Different

so the instances with the shortest execution time matched. On the other hand, the instance able to run the Sysbench prime number process at lowest cost was c4.xlarge, while for Hadoop word count, it was c4.large, so the lowest cost instances did not match. This result appears on the first row of data in Table 7. From left to right, it indicates that for the prime number process in Sysbench, the similar application was Hadoop word count, that the instances with shortest execution time matched, and that the instances with lowest usage cost did not match.

6. Discussion

We now discuss the results of these evaluation experiments. Instances for which execution with short execution time and low usage charges are possible tended to be c4.xlarge and c4.large, respectively. The cause of this is that many of the applications used in this experiment were CPU-bound, and this is considered to have had a major impact on the match rate. In particular, the prime number calculation by Sysbench and the performance required for Dhrystone in UnixBench is biased toward the CPU. As the processing for these had the shortest execution time and lowest usage costs in c4.xlarge, which is optimized for the CPU, this is a result compatible with the published ECU values. For the scientific computing workflow, processing differs depending on the task, and there are tasks that require a lot of non-CPU processing. For that reason, in the reading and writing of memory for Sysbench, which is an application that requires not only CPU, but also memory and I/O processing performance, and file sorting by Hadoop, execution time became shorter due to CPU performance. CPU performance is important even for applications requiring memory and I/O processing performance; therefore, creating a calculation formula that is weighted in consideration of the impact of each on execution time and usage cost is effective when selecting resources.

To apply this method to the scientific computing workflows carrying out a variety of processing, it is necessary to increase the number of sample applications handled and support a more diverse range of processing. Additionally, as the sample applications used in this test have a low computational volume, there is a concern that it cannot support scientific computing workflows handling huge volumes of processing. A greater diversity of sample applications is required to realize this method and enable the selection of computational resources in scientific computing workflows.

7. Conclusion

The objective of this experiment is to achieve a method of selecting resources based on execution time and usage costs when using commercial cloud services for scientific computing workflows. By using instance performance and the execution performance values required for the application, we measured the features of the application for a certain instance. The aim is to propose a method for selecting instances that can be executed in the shortest execution time with the lowest usage costs, by referencing similar applications for the measured execution performance values. In the evaluation experiment, we verified the effectiveness of selecting resources based on similarity. The match rate of the results was approximately 55%, and a large impact was present in CPU-bound applications. As this considers the impact on execution time when making resource selections more than for memory or I/O, it is necessary to focus on CPU performance.

Our experiment showed that the proposed method based on similarity can usually select the best instance for Hadoop or Sysbench-type applications. Furthermore, scientific computing workflow applications are mainly executed using few system functions, like Hadoop and Sysbench-type applications. Thus, our method would be effective in selecting resources for many types of scientific computing workflows. On the other hand, if the application requires many system functions, like UnixBench, the similarity calculation requires weighting factors to handle complicated behavior.

References

[1] T. Koita Performance Evaluation of Memory Usage Costs for Commercial Cloud Services, Proc. of the IEEE 3rd Int'l Conf. on Big Data Analysis (ICBDA2018), pp.307-311, 2018.
[2] Ewa Deelman, Pegasus and DAGMan From Concept to Execution: Mapping Scientific Workflows onto Today's Cyberinfrastructure, Proc. of the Advances in Parallel Computing, vol.16, pp.56-74, 2008.
[3] Open Science Grid, https://www.opensciencegrid.org/.
[4] Amazon Elastic Compute Cloud (EC2), http://www.amazon.com/ec2/.
[5] Y. Manabe, Performance comparison of scientific workflows on EC2, IPSJ technical report, 2016.
[6] S. Ostermann, A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing, Proc. of the Cloud Computing, pp.115–131, 2010.
[7] G. Juve, Scientific workflows and clouds, Crossroads, vol.16, pp.14-18, 2010.
[8] G. Juve, Scientific workflow applications on Amazon EC2, Proc. of the 5th IEEE International Conference on e-Science Workshops, pp.59-66, 2009.
[9] B. Tovar, A Job Sizing Strategy for High-Throughput Scientific Workflows, IEEE Trans. On Parallel and Distributed Systems, vol.29, no.2, pp.240-253, 2018.
[10] I. Sfiligoi, Estimating job runtime for CMS analysis jobs, Proc. of J. Physics: Conf. Series, vol. 513, no. 3, 2014.
[11] Y. Manabe, Resource provisioning method for scientific workflows on commercial cloud services, graduation thesis, Doshisha University, 2017.
[12] Sysbench,http://imysql.com/wp-content/uploads/2014/10/sysbench-manual.pdf, 2009.
[13] UnixBench, http://code.google.com/p/byte-unixbench/.
[14] Hadoop, http://hadoop.apache.org.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 72-81 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Machine Learning Framework Using Distinctive Feature Extraction for Hand Gesture Recognition

Sudipta Saha[*,1], Aninda Saha[2], Zubayr Khalid[1], Pritam Paul[1], Shuvam Biswas[1]

[1]*Computer Science and Engineering, Institute of Engineering and Management, Kolkata, 700064, India*

[2] *Electronics and Communication Engineering, Institute of Engineering and Management, Kolkata, 700064, India*

ARTICLE INFO

Article history:
Received: 18 July, 2018
Accepted: 05 September, 2018
Online: 18 September, 2018

Keywords:
Hand gesture recognition
Feature extraction
Polygon approximation
Convex decomposition
Machine learning

ABSTRACT

There are more than 7billion people in the world where there are around 500 million people in the world who are denied from normal lifestyle due to physical and mental issue. It is completely fair to say that every person deserves to enjoy a normal lifestyle. While physically and mentally challenged people find suitable way to surpass their limits, thus become able in other ways, researchers always try to find solutions better than the existing one. A complete remedial of such issue is included in advanced medical science, and the amelioration of such issue to a better extent is the challenge for the engineers.

In this work we have focused on hand gestures. Hand gestures are created using the movement of hand and arm, using fingers to create different shapes, using fingers and palm to create different angles. Single or both hands can be used to create different expressions. The main objective of this work is to generate an algorithm that can recognize different patterns of hand gestures with notable accuracy. American Sign Language is one possible reference model that can be used. Images of different hand signs are taken as inputs using a webcam, followed by segmentation of the images using polygon approximation and approximate convex decomposition. Feature extraction is done by recording the unique feature among the various convex segments of the hand. The resultant singularities are then used as extracted feature vectors. This involves training with the obtained features which are approximately unique for different hand gestures. Hence, we will be able to identify sign languages and successively make disabled individuals socially acceptable. This work is an extension of the work entitled "A Machine Learning Framework Using Distinctive Feature Extraction for Hand Gesture Recognition" in 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (IEEE-CCWC).

1. Introduction

Gesture Technology is an emerging scope in the field of Computer Science, precisely machine learning. The objective of gesture technology is to interpret or recover the human gesture using intelligent machines. Complex mathematical algorithm is used to interpret the gestures. Fundamentally spoken, the mathematical algorithms contain a large set of test and training data, which are eventually used to compare with the current status of the human, thus come out with an appropriate result that can describe the gesture. Gesture is vastly used in human-machine interaction where the machine detects gesture input from the human and responds accordingly. Advanced technology has created a possible way for communication between 2 or more machines, with no human presence, using gesture input and output. Furthermore, gesture technology can be very useful in human to human interaction, precisely where the persons involved in the communication do not have the same mode of communication. While other body parts are also used in gesture technology, the most used parts are face and hands. In general term, gesture can be divided into 2 types- a) body gesture: where body is used to create expression; and b) hand gesture: where only hand is used to create expression. A sign or symbol is the most commonly used hand gesture [1]. Gesture can be further distinguished on basis of static gesture [2] and dynamic gesture [3]. Face images in gesture technology are mostly used for emotion detection [4] and as a security tool used in gadgets and other devices [5]. Some main application areas for gesture technology include gaming sector, home automation, automotive sector, sign language interpretation.

[*]Sudipta Saha, Kolkata, India, +918013554858, subho040995@gmail.com

In hand gesture technology, human can interact with machine or with another human through a machine that recognizes the hand gestures created by the human being [6]. Generally still images of hand gestures are sent as inputs to the machines, but there are also machines that can collect gesture data from live video footage. Thanks to advanced technology, touching of the machine is not necessary anymore; strong sensors in the machine can detect hand gestures from distance. Hand gesture technology is verily used in the later two of the previous example. In this domain, we will talk about hand gesture technology and sign language interpretation.

Enjoying the proper lifestyle is the birth right of every human being in this world. Everybody should be able to communicate and share expression with the outer world, but unfortunately not every time people are able to do so. It is always hard for people to communicate with others when mental or physical challenges become a barrier in the way. While it is safe to say that these people are verily able in other ways that help them to communicate, we always try to find a system better than the existing one. The most fundamental and most popular technique used in this field is creating sign language using hand gestures [7]. But the challenges occur when at least one of the people involved in such communications doesn't know the sign language or how to create them. A few numbers of people who don't need sign language for communication actually know this language, resulting in a great difficulty when there is a communication between 2 people where one of them only uses sign language for sharing expression. Based on practical experience, it is almost certain that these types of conversations do not proceed easily unless there is a third person who is used as an interpreter. There exist devices, as for example, data gloves, which are used to solve such problems during conversation. Data glove is an interactive device, basically a glove to wear with sensors on it. These sensors create precise capturing of hand gesture information. Though data glove is generally used in robotics, other application areas also include human interaction. This report refers only to camera based static hand gesture recognition [3]. Static hand gesture recognition technology has a useful application on communication between 2 persons where one is able to use a particular sign language (say ASL or American Sign Language) and other isn't. A hand gesture recognition device is used as the protocol in this communication. As mentioned earlier, the device contains sensors to capture images and to detect gesture data and analyze them. The basic objective of the device is to eliminate the necessity of a third party interpreter, by allowing itself to do the same job. It captures gesture images as input and creates output suitable for non-ASL speaker. Generally the output is in textual format. The device is also used in the reverse situation, that is., taking textual or vocal words as inputs and creating sign language as outputs. Since the algorithms used in a device is based on a particular sign language, it is necessary that mode of communication from the sign language user's side only includes one specific language, such as either Indian Sign Language (ISL)[8] or the American Sign Language (ASL)[9] but not an amalgam of both. The objective of this paper is to illustrate the extraction of unique features from each symbol or gesture in reference to an existing data set (ISL or ASL) and then train the machine with the obtained feature vectors using standard classification models.

Using a huge number of data in test set and training set creates less ambiguity for the machine while taking decision, and as a result it can give a more accurate output. As for example, a specific hand sign can be described by several textual words, and a set of hand signs can be described by 2 or more different sentences. The more there is data in the device's memory based on previous trainings (that is, in training data set), the more accurately the device will be able to decode texts into proper sign languages. There are more necessities to be taken care of in order to get more accurate results. Use of white gloves is employed to eliminate unwanted wrinkles or groves on the palm and fingers. From the threshold image of the figure we obtain its minimum fitting polygon through what we call polygon approximation technique. The next step is Convex Decomposition technique. It segregates each of the convex parts of the hand, which is an essential step to identity the crucial aspects of the gesture. The next and final step is to extract the required unique features. The wrist point in the hand is considered to be the mid-point of the lowermost polygon.

The novelty of the proposed method is listed as follows.

- Extraction of individual features for each and every gesture creates the possibility of notable polarity and a chance of further efficient classification of the obtained data.
- The proposed method is equipped for large scale deployment through recognizing the demand and necessity of real time application.
- The produced results from trained dataset are favorable with adequate acceptance rate.

2. Related Work

Both the Indian Sign Language (ISL) and American Sign Language (ASL) are comprised of a notable list of hand gestures, each with meaningful expression that can be used for any non-verbal communication. As mentioned earlier, the sign languages come real handy when we have the technology to decode a specific sign language to sounds/texts and vice-versa, thus useful in communication where one person is accustomed with sign language communication where as other person use vocal/ textual communication and does not know the sign language. In probability, Markov Model is a concept for modelling randomly changing system. According to this model, the next state of a random system mostly depends on the current state of the system; and not on the previous states. In 1960 William Stokoe developed a concept based on real time Markov-model based systems [10]. In this system, continuous American Sign Language (ASL) generates long combination of words as outputs. A single camera is used in this concept. The concept is still very much effective and is majorly used technique to recognize gestures. However, this implementation requires polygon approximation followed by convex decomposition.

The convex decomposition method is a special case of alternating sum of volumes (ASV) method. ASV can be defined as a series of convex elements stick together in a way such that for elements A, B, C, D, E, … the series will be: A u B / C u D / E and so on, where u is Union operation and / is Difference Operation [11]. In 1998, Starner proposed the using of convex hulls and set- difference operations [12]. A convergent convex decomposition is proposed that uses a combination of ASV decomposition and remedial partitioning for the non-convergence.

The concept uses ASVP decomposition for image feature extraction, in other words, to identify the unique properties of an image that differentiates the image from others.

Polygon Approximation is very necessary in this context. The simplest idea of the process can be described by 4 steps – Identifying original shape (contour), chopping the shape around the corners, generate approximate convex lines for the chopped segments, merging the segments [13]. The decomposition of polygon with 0 or more holes into convex chops can be done approximately using Approximate Convex Decomposition (ACD) [14]. The algorithm used in this context is simple: iterative resolving of the most significant non-convex feature. Approximate Convex Decomposition method is applicable for both 2D and 3D models.

On the basis of simple cut and split paradigm of Chazelle, a method to compute a convex decomposition of a non-convex polyhedron of arbitrary genus (handles) and shells (internal voids) is generalized [15]. The algorithm is extended to work for a certain class of no manifold polyhedral.

A set of topological decimation operations to its dual graph is applied to calculate a hierarchical segmentation of mesh triangles, which is used for 3D mesh approximate convex decomposition [16]. The method of segmentation incudes a) portraying a triangle mesh and its corresponding dual graph, b) contracting of an arc of the dual graph, followed by the marking of the end-points of the two triangles corresponding to the dual arc's as belonging to the same single cluster [17]. Segmentation can be of two types – direct segmentation and feature based segmentation. This particular model is important due to its adaptation to collision detection.

Inter-class variability enhancement and local-global ambiguity identification for each hand gesture is used in the automatic gesture recognition on Indian Sign Language (ISL) [18]. In ISL both hands can be used to represent an expression. Support Vector Machines (SVM) is used to classify each hand posture. In SVM data items are plotted as points in n-dimensional space, where n is the number of features. To classify dynamic gestures, Dynamic Time Warping (DTW) is used with the trajectory feature vector, and it achieves 86.3% recognition rate.

A pre-trained Google Net architecture representing a real-time sign language interpreter is described while developing the architecture with ILSVRC2012 dataset as training sets [19]. The idea is based on Convolution Neural Network (CNN). The main difference between Convolution Neural Network (CNN) and ordinary Neural Network is that CNN assumes that all the inputs are images, allowing researchers to encode certain properties into the architecture. The architecture is also trained with data sets from Surrey University and Massey University ASL data. This training set allows the architecture to apply transfer learning to this task. The system features a pipeline that captures video of signing a word as input through a designated application. Individual frames of the video are extracted and then they generate letter probabilities using a CNN.

Initially made use of Histogram of Oriented Gradients method extraction of feature, a feature extraction method uses two neural networks based recognition of 10 gestures [20]. The HOG [21] for an image describes the elementary property (such as color, shape and size) of an image. The objective of HOG method is object detection.

Data availability and analyzing their relative merits to determine features are essential for sign-language recognition (SLR) [22]. The objective is to continuous sign recognition and design a model that is independent of signer. The model should also be adapting to larger and noisy data sets. A statistical analysis based on four parameters: position, posture, orientation, and motion is used for a large vocabulary sign language interpreter with real-time continuous gesture recognition of sign language [23].

3. Data Set

The dataset for the proposed method has been centered on the popular and ubiquitous ASL (American Sign Language). It was formulated by the University of Surrey's Center for Vision, Speech and Signal. The images referred have been indicated in the figure (1). We were able to gather approximately more than 500 colored images however the algorithm is constrained to static gesture recognition.

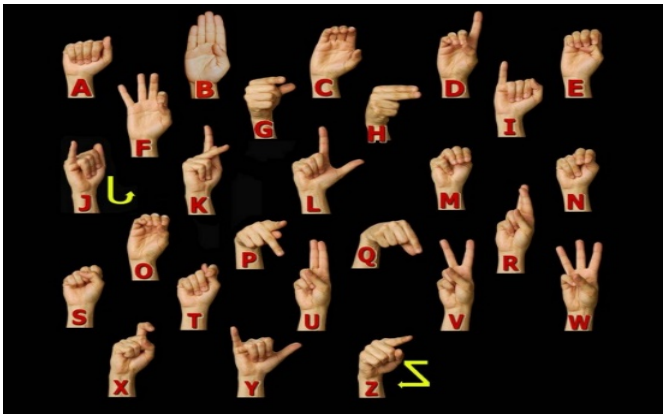


Figure 1: Different Gestures of ASL

4. Methodology

4.1. Pre-Processing

In the first step where we remove uneven parts of the image, dilate and erode and then generate a contour based on color edge detection. In the above set of figures, figure (2.a.I) is the initial figure, which is followed by figure (2.a.II) where we get the processed image. Using the processed image, we can find out the contour of the dark area using color edge detection technique [24] shown in the figure (2.a.III).

Once the contour is obtained then the rest of the area apart from the contour can be removed. We assumed that the hand will fit to (having minimal space between the hand and the frame border) the frame with the frame dimensions being constant. Once the background is removed, we need figure out the edges of the fingers and hands marked by black contrasting borders. Figure (2.a.IV) shows the finally obtained contour after preprocessing.

4.2. Approximation of Minimal Fitting Polygon

In digital image processing the approximation of arbitrary two-dimensional curves by polygonal figures is an imperative technique. In number applications, this technique is widely used

for a purpose to represent lines and boundaries by means of polygons that have a minimum number of vertices so that fit condition is satisfied. Using island finding algorithm, we will be able to separate the fingers by inner smaller polygons.

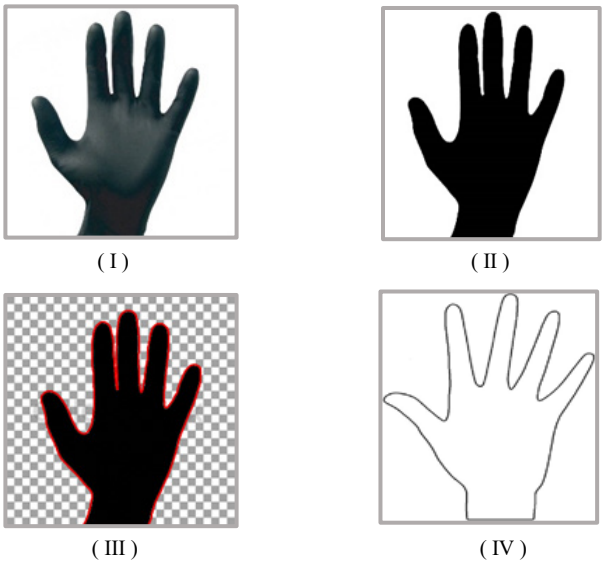


Figure 2 (a): Images Describing the Pre-Processing stage

To find the initial point we need to move by traversing alone the x-axis of the bottom frame and then we can find the first pixel that is colored black. The adjacent pixel can be located by a technique where we check the neighboring 8 pixels for a black pixel. The algorithm navigates through the contour of the figure from pixel to pixel, marking them based on their connectivity and the relative values of their neighbors. Connection between the pixels can be determined with the help of a 3x3 grid as shown in Figure (2.b). This grid can be either 4-connected or 8-connected.

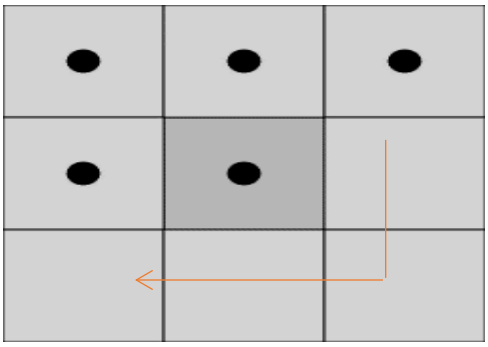


Figure 2 (b): 8 neighboring pixels of a pre-determined pixel

Now we need to find the 2D polygon by the following algorithm:

1. Now take a black pixel at the starting point and continue moving along the pixel.
2. Now joining the starting point pixel and adjacent point pixel (pivotal point) with a line to forms the base. Now we calculate the angle between the base and the next black point and store the calculated angle.
3. Now if calculated angle is not considerably larger or smaller than the previously calculated angle, then we need to eliminate the current pixel point, and take the next pixel point lying in succession as the pivotal point. Else, if the angle is substantially more or less by a threshold value, then consider the current point and initialize the point as starting point. Go to step

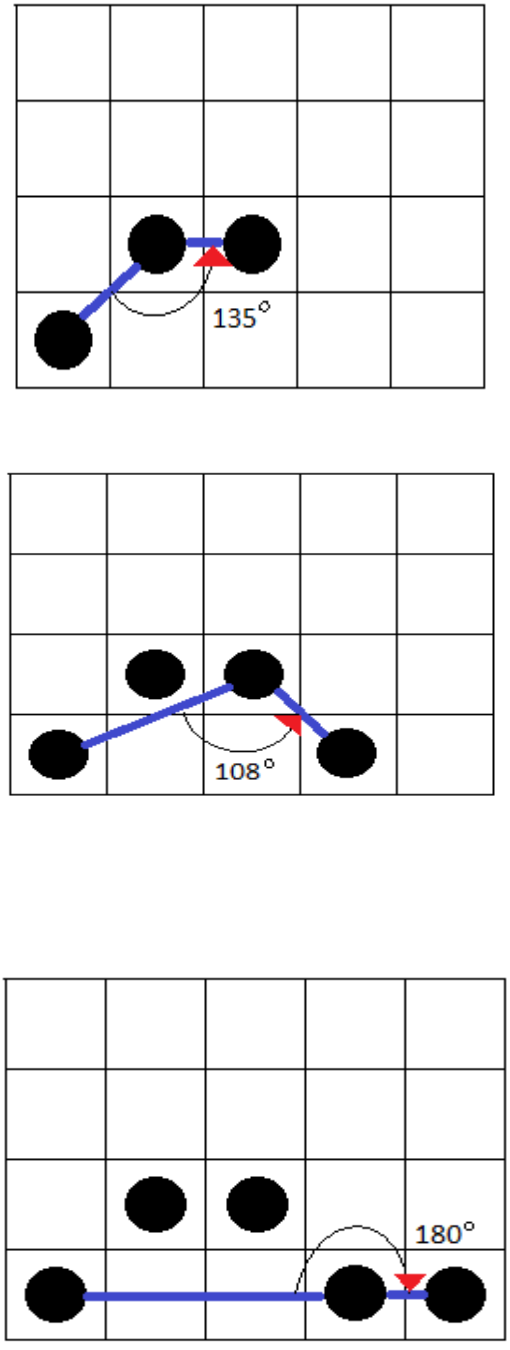


Figure 3 (c): Since 180° is considerably large, it is accepted

4.3. Island Finding Algorithm:

Here, we shall try to find out the minimal cyclic paths that enclose an area. A DFS traversal is done using the following constraints

1. Stack(S) ← Push(R); S is the stack and R is root node.
2. If S==null, end looping, else,

3. If current node X present in Stack S, go to step 6.
4. Else if X not present in Stack S and if it has child nodes then expand and explore child nodes if [Child(X)!=visited or Child(X)!=Parent(X)]. Else return to parent node.
5. Repeat step 2.
6. If substring[X, indexOfLastNode] or reverse (substring[X, indexOfLastNode]) is not already stored, then store (substring[X, indexOfLastNode] + X) as cyclic path C_n. Mark X as visited, POP(X) and PEEP(S). Go to step 5.
7. If substring [X, indexOfLastNode] or reverse (substring [X, indexOfLastNode]) is already stored, then mark X as visited, POP(X) and PEEP(S). Go to step 5.

Once we find all the cyclic paths possible (C_1, C_2, C_3….C_N) then in a sequential order consider each cyclic paths, such that if elements of (C_n _ lastElement) or elements of reverse (C_n _ lastElement) is subset of C_m, then discard C_m. Continue until we have minimal unique cyclic paths possible. Then a given cyclic path forms either an island (finger or palm) or it's the outer contour of the figure.

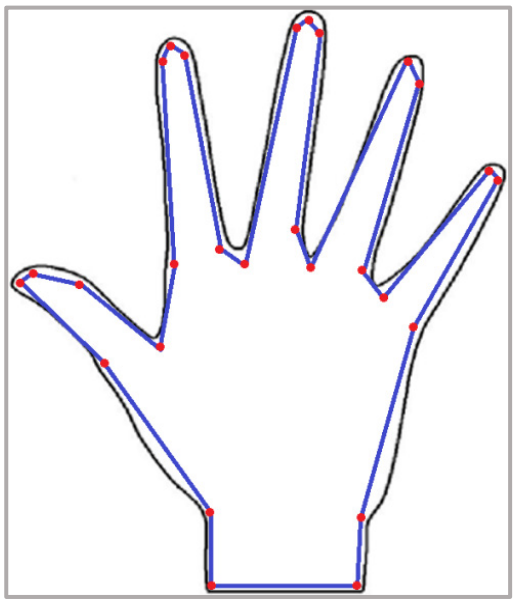


Figure 4: Illustration of Polygon Approximation of Polygon from the Shape of a Hand

In the above figure 4, the polygon with vertices of color red is sketched out from the curved outline of the image. A simple polygon P is formed when the vertices are joined by the blue lines.

The non-essential curves and bends are removed by the technique of approximation of minimal fitting polygon [25] and give us a figure that is bounded by line segments. The resultant image consists of a simple polygonal region in the plane bounded by a non-self-intersecting, closed, polygonal path. The polygonal path itself is part of the polygon; it is usually called the boundary.

4.4. Decomposition into Convex Components

Now if we partition the complex models into pieces it will be easy to deal with complex models or many-sided figure. Numerous difficulties can be solved more proficiently when the input is convex. Convex components Decomposition into results in pieces that are easy to process and additional approximation simples and eases computation, ensuring quicker and efficient output with manageable margins of inaccuracy.

Exact and approximate are two type of convex decomposition. Number of clusters in the model should be minimized so we do not use exact decomposition. We can get minimum number of clusters using Approximate convex decomposition ensuring that each cluster has a concavity lower that a predefined threshold.

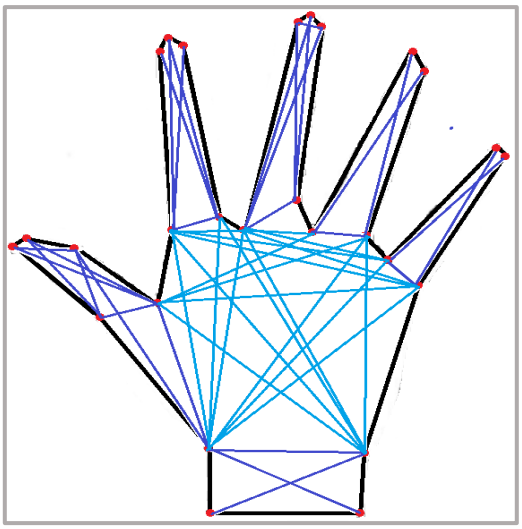


Figure 5: Polygon Decomposition into Non-Concave Sub graphs

The figure 5 given above shows the result of convex decomposition on the approximated polygonal segments obtained in the previous section.

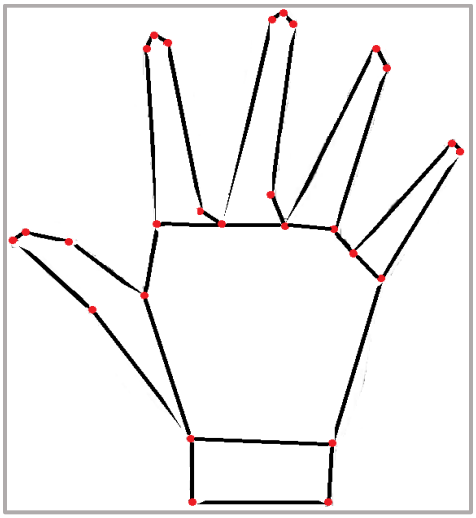


Figure 6: Resulting Convex Decomposition

In graph theory, where a complete subset of vertices of a graph G such that every two distinct vertices in the clique are adjacent is known as a clique. Subsequently, a maximum clique is one that has the largest possible number of vertices among other cliques. We then observe the maximum clique possible for any given point and then iteratively check for the points it connects. For given vertices V = {V_1, V_2, V_3…V_N},the we need to form two groups such that if $V_M \in V$ then, put VM into Checking array(C), and the vertices connecting V_M in Pointing array(P). Put P's last

element into C and iteratively check if C elements has common connections with vertices in P. Group those points to find if they share any common point to find the approximate convex partition [26]. The maximum clique in the above mentioned visibility graph G, which is also undirected, is a complete sub graph of G having the maximum number of vertices.

4.5. Detecting Wrist Point and Reference Line

Deriving the data necessary for feature extraction the wrist point and the reference line are needed. The data are calculated with respect to the reference line. As the image is always considered to have the picture of the gesture in upright position so the wrist point is calculated as the mid-point of the bottom most polygon.

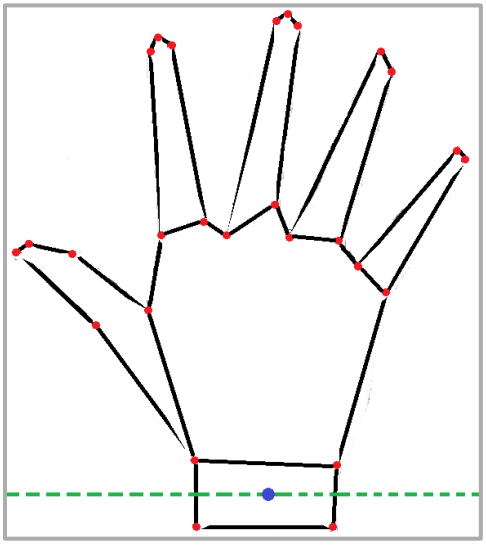


Figure 7: Deduction of Wrist Point and Reference Line

First identify the polygon associated with the wrist to derive the reference line. Then the mid-point of the polygon is calculated. The line that passes though the wrist point and is parallel to the bottom of the frame of reference is known as reference line. The green dotted line in figure 7 indicates the reference line. The bottom-most polygon is the polygon associated with the wrist, and the blue dot is the midpoint associated with the polygon.

4.6. Angle and Distance calculation with respect to Wrist Point

We now have the line of reference as well as the wrist point which are used to realize the different polygon segments in a given figure. To achieve this by observing the incrementing angles between each line joining the mid points of the polygonal segments and the line of reference. Our main objective here is to find unique feature vector to represent different hand gestures.

With reference to figure 8 by the convex decomposition process we begin by finding the mid-points of the segments. These points, marked as P1 through to P6, are then extended to join the wrist point marked as W. What is important to note here is that with each segment mid-point, the angle increases with the minimum being a possible 0° and the maximum being a possible 180°.This gradual expanding gap between the wrist point and each subsequent line signifies the change in finger positions and helps us to differentiate between them. Another aspect of the line joining the wrist point and the mid points of the segments is its length.

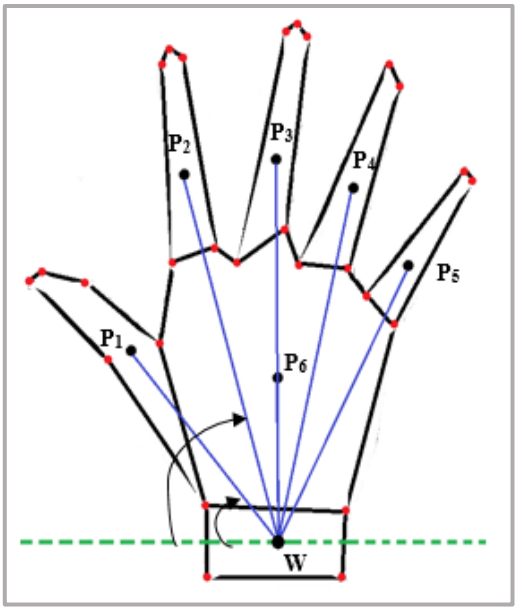


Figure 8: Connection of Wrist Point, described by P1, P2, with the Centroid (W) of the Decomposed Polygons

The property which helps us to decide whether a finger is in a folded or unfolded position is the distance between the wrist point and the mid-point of the corresponding polygonal segment of that finger. While the finger is in a folded position, the centroid of the segment gets located in a vertically lower level inside the polygon than it would if the finger was in an unfolded position. As a result, it becomes fairly easily distinguishable to recognize different signs, especially the ones that only differ by the positioning of a single finger.

Outcome of the above stated procedure produces the angle and distance of each mid-point of the different generated convex segments with respect to the reference line and wrist point. It shall be noted that the duality (angle and distance) needs to be converted to a singularity for feature extraction and forming a unique one-dimensional array representation of each distinguishable hand gesture.

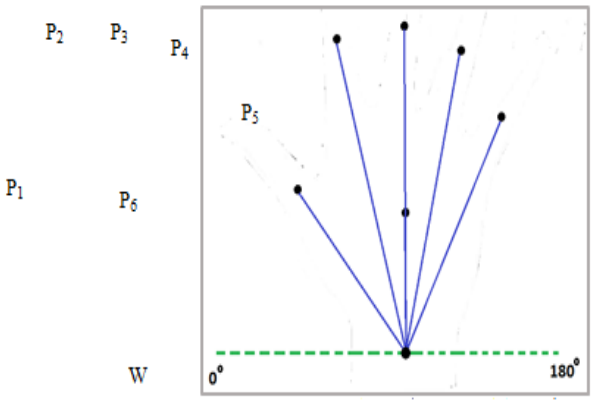


Figure 9: Unique Array of Data to Represent a Hand Gesture

pred - NumPy array

	0	1	2	3	4	5	6	7	8	9
0	0.866076	0.0158627	0.000789759	1.44649e-05	1.95762e-06	3.93198e-05	1.31903e-08	5.32135e-06	9.7464e-08	0.117211
1	0.979508	0.000120881	0.0179403	7.42718e-10	1.49179e-08	1.10229e-07	0.00242796	2.21401e-06	2.99634e-07	1.00516e-10
2	1.64009e-05	0.981988	2.73922e-10	3.00392e-07	6.03354e-06	0.0179857	2.21962e-06	8.16551e-07	8.16551e-07	2.02403e-09
3	1.41648e-08	0.930058	2.10224e-06	0.0170346	0.0463049	1.91699e-09	5.71448e-06	0.000312	1.55336e-05	0.00626668
4	0.00235374	0.047276	0.949565	2.90474e-07	0.000318543	0.000318543	2.14633e-06	4.31102e-05	0.000117186	5.83433e-06
5	5.84761e-06	1.44948e-08	0.951727	1.96165e-09	0.0473837	1.58955e-05	2.65481e-10	5.33233e-09	3.94009e-08	0.000867862
6	1.48066e-08	5.97342e-06	0.00240985	0.972203	0.000119979	1.62375e-05	8.08415e-07	0.000886535	0.00655066	0.0178065
7	0.000328505	2.21345e-06	0.00242734	0.979259	0.0179358	7.42529e-10	4.44583e-05	7.42529e-10	4.05407e-08	2.21345e-06
8	8.14315e-07	2.21354e-06	5.48681e-09	0.0179365	0.979298	0.00242744	6.01702e-06	2.9957e-07	4.05423e-08	0.000328518
9	1.06265e-07	1.43814e-08	2.13438e-06	0.0470129	0.944279	0.0063625	0.00234063	2.88857e-07	2.13438e-06	1.06265e-07
10	0.000892449	6.01328e-06	2.73002e-10	0.0179253	1.63458e-05	0.978689	0.00242593	4.44324e-05	4.05171e-08	4.05171e-08
11	0.00626461	0.000311897	4.22106e-05	1.41601e-08	0.000311897	0.929751	0.017029	0.0462896	5.2092e-09	1.0463e-07
12	0.000121066	2.2174e-06	3.00093e-07	0.0179678	2.73649e-10	7.43855e-10	0.981008	0.000894563	6.02752e-06	4.06131e-08
13	3.91773e-08	1.58052e-05	2.13901e-06	0.000116786	1.06495e-07	0.0471148	0.946325	5.81442e-06	0.00637629	4.29631e-05
14	7.421e-10	1.00432e-10	0.00242594	0.000892452	2.21217e-06	4.44326e-05	1.63458e-05	0.978693	2.99384e-07	0.0179254
15	2.98352e-07	5.46451e-09	0.0178636	5.46451e-09	5.99256e-06	0.000120364	0.000120364	0.975318	4.03775e-08	0.00657164
16	0.0178459	1.48394e-08	1.62734e-05	8.10204e-07	4.03377e-08	0.000888496	0.000326859	0.00656515	0.974354	2.20236e-06
17	0.0178169	2.71351e-10	0.00241126	2.00503e-09	4.02721e-08	5.45024e-09	0.00655448	0.000120049	0.972771	0.000326328
18	1.10398e-07	0.0179678	0.000894563	6.02752e-06	1.10398e-07	2.2174e-06	8.15736e-07	0.000121066	2.02201e-09	0.981007

pred - NumPy array

	0	1	2	3	4	5	6	7	8	9
0	True	False	False	False	False	False	False	False	False	False
1	True	False	False	False	False	False	False	False	False	False
2	False	True	False	False	False	False	False	False	False	False
3	False	True	False	False	False	False	False	False	False	False
4	False	False	True	False	False	False	False	False	False	False
5	False	False	True	False	False	False	False	False	False	False
6	False	False	False	True	False	False	False	False	False	False
7	False	False	False	True	False	False	False	False	False	False
8	False	False	False	False	True	False	False	False	False	False
9	False	False	False	False	True	False	False	False	False	False
10	False	False	False	False	False	True	False	False	False	False
11	False	False	False	False	False	True	False	False	False	False
12	False	False	False	False	False	False	True	False	False	False
13	False	False	False	False	False	False	True	False	False	False
14	False	False	False	False	False	False	False	True	False	False
15	False	False	False	False	False	False	False	True	False	False
16	False	False	False	False	False	False	False	False	True	False
17	False	False	False	False	False	False	False	False	True	False
18	False	False	False	False	False	False	False	False	False	True

pred - NumPy array

	0	1	2	3	4	5	6	7	8	9
0	0.951614	5.3317e-09	0.0473781	2.91101e-07	7.21567e-10	5.84692e-06	0.00086776	7.91294e-07	0.000117438	1.58936e-05
1	0.978299	1.48995e-08	0.0179182	0.000328183	5.48121e-09	1.10093e-07	0.000120732	1.63393e-05	0.000892094	0.00242496
2	1.44711e-08	0.950175	0.00235525	4.31379e-05	0.0473064	7.90098e-07	0.000117261	7.20476e-10	1.44711e-08	2.14771e-06
3	1.12529e-07	0.999945	6.14388e-06	2.26021e-06	4.53974e-05	1.52291e-08	8.31483e-07	2.78932e-10	2.06104e-09	4.13971e-08
4	2.02399e-09	4.45813e-05	0.98197	1.10506e-07	1.49554e-08	0.0179854	4.06529e-08	3.00387e-07	7.44584e-10	5.50178e-09
5	0.000852723	0.0465571	0.935125	0.000313699	2.11369e-06	0.0171274	5.7456e-06	1.56182e-05	7.77583e-07	1.05234e-07
6	1.34973e-05	2.47211e-07	4.96537e-06	0.808137	0.0402348	0.00200317	0.109369	2.47211e-07	1.82666e-06	0.0402348
7	5.44774e-09	1.62395e-05	0.000886645	0.972325	0.0178087	0.00241015	7.37271e-10	2.19777e-06	2.00411e-09	0.00655147
8	0.0449461	0.000302844	5.54679e-06	0.000823216	0.902766	0.00608279	7.50676e-07	1.50777e-05	0.0449461	0.00011141
9	0.00657231	0.000120376	8.11087e-07	2.20476e-06	0.975417	2.01048e-09	5.99317e-06	1.62911e-05	0.0178654	1.00096e-10
10	0.017809	1.48087e-08	5.97424e-06	0.00655155	0.000886655	0.972336	1.09422e-07	2.7123e-10	0.00241018	8.08525e-07
11	2.99303e-07	1.00405e-10	8.1359e-07	0.00089221	0.000328226	0.978427	6.01166e-06	0.0179205	0.00242528	1.10107e-07
12	0.000121022	2.21659e-06	4.45213e-05	0.000894235	1.10357e-07	5.49437e-09	0.980647	0.0179612	0.000328971	2.99982e-07
13	1.5536e-05	5.71536e-06	0.046312	2.10256e-06	7.7349e-07	0.000114796	0.930202	4.22311e-05	0.00626765	0.0170372
14	0.0178248	5.97954e-06	0.00655736	1.09519e-07	2.00591e-09	4.02898e-08	8.09243e-07	0.973199	0.00241232	9.98685e-11
15	3.0009e-07	1.49406e-08	0.0179676	1.00669e-10	0.000894556	0.000121065	3.0009e-07	0.981	2.73647e-10	1.63844e-05
16	0.000328985	0.017962	0.000894274	1.10362e-07	0.000121027	2.73561e-10	2.21668e-06	8.15472e-07	0.980691	4.06e-08
17	1.09744e-07	5.46381e-09	2.20426e-06	1.48522e-08	0.000327141	8.10901e-07	0.0178613	0.0065708	0.975193	4.42737e-05
18	8.16502e-07	2.73906e-10	2.21948e-06	6.03318e-06	0.0179847	2.02391e-09	1.63999e-05	1.63999e-05	4.45795e-05	0.981929

pred - NumPy array

	0	1	2	3	4	5	6	7	8	9
0	True	False	False	False	False	False	False	False	False	False
1	True	False	False	False	False	False	False	False	False	False
2	False	True	False	False	False	False	False	False	False	False
3	False	True	False	False	False	False	False	False	False	False
4	False	False	True	False	False	False	False	False	False	False
5	False	False	False	True	False	False	False	False	False	False
6	False	False	False	True	False	False	False	False	False	False
7	False	False	False	True	False	False	False	False	False	False
8	False	False	False	False	False	True	False	False	False	False
9	False	False	False	False	True	False	False	False	False	False
10	False	False	False	False	False	True	False	False	False	False
11	False	False	False	False	False	True	False	False	False	False
12	False	False	False	False	False	False	True	False	False	False
13	False	False	False	False	False	False	True	False	False	False
14	False	False	False	False	False	False	False	True	False	False
15	False	False	False	False	False	False	False	True	False	False
16	False	False	False	False	False	False	False	False	True	False
17	False	False	False	False	False	False	False	False	True	False
18	False	False	False	False	False	False	False	False	False	True

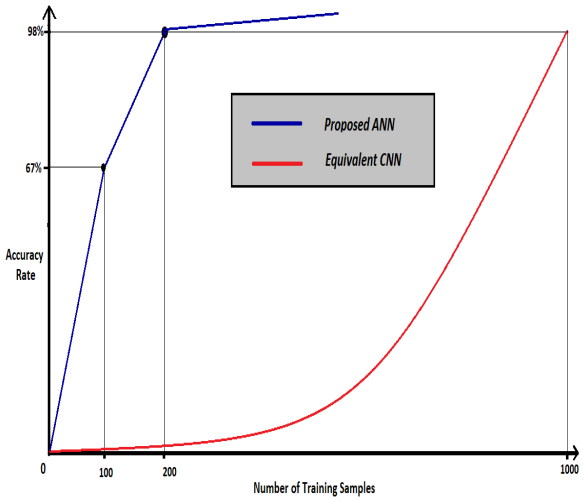


Figure 11: Graphical Explanation of Accuracy versus Number of examples.

To generate a feature vector, convert the duality (angle and distance) into a singularity is the main challenge. For a given angle and length use a preexisting formula, ***lsinΘ*** or the sin component. The following table is a sample of generated feature vector for the data obtained from table 1.

Table 1: Calculating the Elements of Feature Vector

	Length(L)	sin(θ)	Lsin(θ)
P1	291.51	0.779	227.85
P2	451.79	0.967	436.79
P3	454.00	0.999	453.54
P4	195.01	1.000	195.01
P4	431.14	0.976	420.75
P5	366.03	0.895	327.59

There are cases where we have found that the number of data points are unequal. As a result, the maximum possible number of data points in a gesture from a given set of gestures is considered as the length of the feature vector array. In case of any gesture having less number of data points than the array size, the empty spaces are filled with 0. This is not analogous to padding with 0, because here it means that the figure has more fingers which are at an angle of 0° or 180°. The feature vector for any gesture is highly distinguished from another gesture.

Once ample number of feature vectors have been obtained, then we can proceed to create our feature file for training the machine. The matrix of features can be created using the feature vectors from the training set. We then train the machine with our matrix of features using Artificial Neural Network and compare the accuracy of each on a common test set.

Data obtained from ten images each of ten alphabets are used for training and testing. Learning is done using ANN classifier. For training and testing we have split the data in 8:2 ratio for each alphabet. Our training set consisted of 10 pictures each of 10 alpha-numeric gestures. Thus we have used 80 images in total to train the machine and further 20 to test the accuracy. Our training included the deployment of cross-validation technique to obtain more precise results. Cross-validation randomizes the data used for training and testing respectively.

5. Results

The aforementioned algorithm has been put to test and has yielded an average accuracy of 92.5%. In reference to the results generated by the code, the confusion matrix of a couple of outputs have been mentioned with resultant prediction values.

In the above stated figures, test 1 has revealed an accuracy of 100% stating that all images of the hand gesture has been successfully classified. The confusion matrix indicates true value for the hand gestures it has predicted for a given feature vector and false for the rest. In the second test the algorithm has missed classified two figures and hence it can be noticed in the confusion matrix where the algorithm has wrongly predicted true for hand gestures numbered 2 and 4 respectively. Additional tests have been conducted in order to obtain consolidate the rate of accuracy.

Table 2: A Comparison between the Proposed Frameworks with the Available CNN

	Proposed ANN	Equivalent CNN
Training Time	3 minutes for 100 sample	18 minutes for 100 samples
Prediction Time	1 millisecond	10 milliseconds
Feature Vector Length	10	255 X 255

The above figures strengthen the advantages of the proposed algorithm over existing CNN or Convolution Neural Network algorithms. Sample training shows that on an average 3 minutes are required for training our algorithm in comparison to CNN which takes roughly 18 minutes for the same number of images. As stated before, storage requirements are highly minimized due to the miniscule feature vector size and the number of images required to obtain adequate accuracy. Thus, for a considerably lesser number of images the algorithm generates acceptable accuracy in contrast to existing algorithms.

Thus the aforementioned framework is readily available for application on computation constricted environment in public sectors and also in absence of high data processing infrastructures.

An average accuracy of 92.5 % has been secured by our algorithm while observing each cycle of training and testing set obtained using cross-validation technique. In comparison to existing Convoluted Neural Network models (CNN), it requires substantially less data and produces acceptable accuracy. Further CNN is computationally expensive model of prediction in comparison to ANN model coupled with distinctive feature extraction.

6. Conclusion

Our technology and technique for gesture recognition has provided significant improvements over previously worked on methods in terms of accuracy and reliability. The disabilities experienced by individuals who are incapable of speaking are overcome with the help of this machine which eradicates the need of another individual for translation of the sign language to the commonly used English language. We have trained our system

with considerably less and manageable sized dataset consisting of various gestures of every letter in the English alphabet. There are some drawbacks, however, in the fact that the process of recognition of a gesture must be carried out with the assistance of a glove without which the creases hinder with the accurate depiction of the hand model and consequently results in imperfect polygon approximation. Also, the process must be carried out in front of a uni-coloured background to ensure that the outer contours of the hand are well defined. But with better and elevated image preprocessing techniques the need for gloves and background may be eradicated. We hope that our research will help the unfortunate ones become socially acceptable and can express themselves spontaneously and freely.

References

[1] Y. Xu,Y. Dai, “Review of Hand Gesture Recognition and Application”; Contemporary Engineering Sciences, Vol. 10, 2017.

[2] T. Messer, "Static hand gesture recognition." University of Fribourg, Switzerland (2009).

[3] J. S. Kim, W. Jang, Z. Bien, "A dynamic gesture recognition system for the Korean sign language (KSL)." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 26.2 (1996).

[4] M. H. Kim, Y. H. Joo, J. B. Park, “Emotion Detection Algorithm Using Frontal Face Image”; ICCAS2005, June 2-5, KINTEX, Gyeonggi-Do, Korea.

[5] D. A. Chowdhry, A. Hussain, Rehman, Z. U. R. Muhammad, “Smart Security System For Sensitive Area Using Face Recognition”; 2013 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (CSUDET).

[6] C. A. Pickering, K. J. Burnham, M. J. Richardson, “A Research Study of Hand Gesture Recognition Technologies and Applications for Human Vehicle Interaction”, 2007 3rd Institution of Engineering and Technology Conference on Automotive Electronics, 28-29 June 2007, Warwick, UK.

[7] P. Pandey, V. Jain, “Hand Gesture Recognition for Sign Language Recognition: A Review”; International Journal of Science, Engineering and Technology Research (IJSETR), Volume- 4, Issue- 3, March 2015.

[8] W. P. Clark, “The Indian Sign Language: With Brief Explanatory Notes of the Gestures Taught Deaf-mutes in Our Institutions for Their Instruction and a Description of Some of the Peculiar Laws, Customs, Myths, Superstitions, Ways of Living, Code of Peace and War Signals of Our Aborigines”, Philadelphia, Pa.: LR Hamersly & Company, 1884.

[9] W. C. Stokoe, "Sign language structure." (1978).

[10] V. Sumalatha, R. Santhi, “A Study on Hidden Markov Model (HMM)”, International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 11, November 2014.

[11] K.Tang, T.Woo, “Algorithmic aspects of alternating sum of volumes. Part 2: Nonconvergence and its remedy”,

[12] T. Starner, J. Weaver, A. Pentland, "Real-time american IEEE Transactions on Pattern Analysis and Machine Intelligence 20.12 (1998): 1371-1375.

[13] K. Selvakumar, “A Review on Polygonal Approximation Techniques of the Digital Planar Curves”, International Journal of Pure and Applied Mathematics, Volume 117 No. 17 2017, 85-90 ISSN: 1311-8080 (printed version); ISSN: 1314-3395.

[14] Y.S.Kim, "Recognition of form features using convex decomposition", Computer-Aided Design 24.9 (1992): 461-476.

[15] J. M., Lien, N. M. Amato, "Approximate convex decomposition of polygons", Proceedings Of The Twentieth Annual Symposium On Computational Geometry, ACM, 2004.

[16] C.J. Bajaj, T. K. Dey, "Convex Decomposition of Polyhedra And Robustness", SIAM Journal on Computing 21.2 (1992): 339-364.

[17] M. Attene, B. Falcidieno, M. Spagnuolo,“Hierarchical Mesh Segmentation Based On Fitting Primitives”, The Visual Computer. 22. 181-193. 10.1007/s00371-006-0375-x, (2006).

[18] K.Mamou, G. Faouzi, "A Simple And Efficient Approach For 3D Mesh Approximate Convex Decomposition", Image Processing (ICIP), 2009 16th IEEE International Conference on Image Processing, IEEE, 2009.

[19] J. Rekha, J. Bhattacharya, S. Majumder, "Shape, Texture And Local Movement Hand Gesture Features For Indian Sign Language Recognition", Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference on Trendz in Information Science and Computing, IEEE, 2011.

[20] M. R. Ahsan, M. I. Ibrahimy, O. O. Khalifa, "Electromygraphy (EMG) Signal Based Hand Gesture Recognition Using Artificial Neural Network (ANN)", 2011 4th International Conference on Mechatronics (ICOM), Kuala Lumpur, 2011, pp. 1-6. doi: 10.1109/ICOM.2011.5937135.

[21] Li, Yuqian, Su, Guangda, “Simplified histograms of oriented gradient features extraction algorithm for the hardware implementation”, 2015 International Conference on Computers, Communications, and Systems (ICCCS) IEEE Xplorer, 2015.

[22] Z. Ren, J. Yuan, J. Meng, Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor", IEEE transactions on multimedia 15.5 (2013): 1110- 1120.

[23] H. Cooper, B. Holt, R. Bowden, "Sign language recognition", Visual Analysis of Humans, Springer, London, 2011. 539-562.

[24] Plataniotis, K.N.; Venetsanopoulos, A.N.; “Color Edge Detection. In: Color Image Processing and Applications”, Digital Signal Processing, Springer, Berlin, Heidelberg, 2000.

[25] E.K. Burke, R.S.R. Hellier, G. Kendall, G. Whitwell , “Complete And Robust No-Fit Polygon Generation For The Irregular Stock Cutting Problem”, European Journal of Operational Research, 179 (2007) 27–49.

[26] P. Prosser, "Exact algorithms for maximum clique: A computational study." Algorithms 5.4 (2012): 545-587.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 72-82 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Interactive Virtual Reality Educational Application

Shouq. Al Awadhi, Noor. Al Habib, Dalal Al-Murad, Fajer Al deei, Mariam Al Houti, Taha Beyrouthy, Samer Al-Kork*

American University of Middle East, Egaila, Kuwait

ARTICLE INFO

Article history:
Received: 19 May, 2018
Accepted: 09 July, 2018
Online: 11 July, 2018

Keywords:
Computer software and applications
Engineering
Information Technology

ABSTRACT

Virtual Reality (VR) technology has become one of the most advanced techniques that is used currently in many fields. The role of education is extremely important in every society; therefore, it should always be updated to be in line with new technologies and lifestyles. Applying technology in education enhances the way of teaching and learning. This paper clarifies a virtual reality application for educational resolutions. The application demonstrates a virtual educational environment that is seen through a Virtual Reality headset, and it is controlled by a motion controller. It allows the user to perform scientific experiments, attend online live 360° lectures, watch pre-recorded lectures, have a campus tour, and visit informative labs virtually. The application helps to overcome many educational issues including hazardous experiments, lack of equipment, and limited mobility of students with special needs.

1. Introduction

"This paper is an extension of work originally presented in BioSMART, the 2nd International Conference on Bio-engineering for Smart Technologies" titled 'Virtual Reality Application for Interactive and Informative Learning' [1]. Significantly, education has been improving throughout the past years by involving technology in this field. In addition, teaching and learning methods are being enhanced by using software and high-tech devices which are making education more effective. Assuming that, every new technology gets linked to education can solve at least one of the problems that face the field especially the students and the teachers. However, a lot of problems are still hindering students and teachers in schools and universities. Distractions and loss of attention during the lecture due to many factors are some of the problems that face the students, and that may lower their educational skills and scores. Furthermore, students are taking the lessons with lack of interest and excitement in the classroom. It is also difficult to visualize the explication of the lessons for some students [2]. Nevertheless, hazardous mistakes, which require practical implementation, are done in labs. For example, some chemical interactions and electrical experiments mistakes lead to considerable serious injuries.

Another problem is, some schools and universities cannot afford some expensive experimental materials. Also, it is hard to provide some biological organs all the time in schools. Finally, the lack of resources plays a huge role in decreasing the educational level of students with disabilities because some of them find it difficult to interact and perform experiments easily due to their special needs and the movements struggles that they face. Therefore, with the purpose of eliminating all these educational matters, technology must be involved with education essentially to build a better sophisticated and educative society.

2. Objectives

Virtual Reality (VR) is defined as "a computer with software that can generate realistic images and sounds in a real environment and enable the user to interact with this environment" [3]. Developing Virtual Reality applications is benefiting education in many aspects. Creating a virtual educational environment can help with solving a lot of educational problems. To be able to have this virtual educational environment different devices have to be connected together. Those devices are Virtual Reality headset, gesture controller or motion sensor, and 360° camera. For the VR headset, it will allow the user to be a part of the virtual world and see the virtual environment. The gesture controller will help him to interact inside this virtual environment. However, the 360° camera will take a realistic 360° images and videos, which will be

*Samer Al Kork, American University of the Middle East, +965 2225 1400 Ext. 1732, samer.alkork@aum.edu.kw

displayed inside the VR environment. In the other hand, many objectives are obtained by developing an educational virtual reality application. Assisting students in comprehending lectures easily and attracting students' attentions by using virtual visualized clarifications of the lecture are some of those objectives. Moreover, the application gives instructors a huge benefit in delivering information easily to the students by supporting their ideas with virtual visualizations. It helps also in performing hazardous experiments without being in danger since the experiments are going to be performed virtually, or student can practice those experiment virtually to be more cautious when they perform it in real-life. In addition, it can solve lack of equipment in labs and affording luxurious equipment problems because the equipment can be provided virtually. Finally, integrating the VR headset with the motion controller will give students with special needs the chance to perform experiments with less movements struggles.

3. Background and Literature Review

Since 1950s, virtual reality was on the boundary of technology but with lack of achievements [4]. However, there was a huge improvement in the virtual reality field in 2012 [5]. The VR headsets are considered as the main tool that enters the user to the virtual world. They are designed to able the user to see the virtual world through special lenses. Also, having a three-dimensional environment is required to be displayed on those VR headset. There are some useful software which help with creating those virtual three-dimensional environments. Those environments can be used as applications that allow the users to learn, watch videos and pictures, and play games.

The literature review discusses the required hardware and software to create a virtual environment and interact with it. First, the hardware part is divided into gesture sensors, VR headsets, and 360º cameras. Second, the software part compares some common software that are used in creating virtual environments and building three-dimensional (3D) applications. Lastly, some existing VR application were mentioned with their purposes.

3.1. Hardware

3.1.1. Gesture Sensors

Gesture controlling technology is a way that is used to detect body motion and gesture. This technology will enable users to interact with some objects and devices without physically touching them [6]. There are many kinds of gesture controllers that can work with the VR technology.

Myo armband is a gesture controller that controls arms' motion [7]. Moreover, Microsoft has produced a sensor controller called Kinect. The motion controller Kinect can be fixed separately on any object like a desk for example, and it is used mostly for gaming [8]. On the other hand, Oculus VR, which is an American technology company, produces touch controllers that are named Oculus touch that are consisting of two controllers that have buttons to control motion by thumbs [9].

Another controller is Leap Motion which is a small device that helps the user to interact with the objects using their free hands. This controller tracks fingers and hands by using three sensors. It can be connected to the PC through a USB port. In addition, it has its own Software Development Kits (SDKs) that helps in programming [10].

3.1.2. Headsets

VR headsets allow the users to enter through a Three-dimensional sphere, so they can look at, move around, and interact with the 3D virtual environments as if they are real environments.

As an example of the VR headsets, HTC Vive headset is considered as a complete set that contains a headset, two motion controllers, and two base stations for defining a whole-room area. HTC Vive has a unique tracking system that tracks the users' movements in 10-foot cube from his position. Also, it provides SDKs for the VR devices in order to simplify the programming process for developers. In addition, it has 32 sensors for 360º motion tracking and a front facing camera which helps in making the virtual world more realistic. The motion tracking system of HTC Vive has two controllers which come with two wireless Lighthouses cameras that senses the signals coming from the headset and the controller's sensors. Then, it programs these signals to make the users moves and interacts virtually in the virtual environment [11].

On the other hand, Samsung Gear VR is considered as one of the most affordable VR headsets. However, it needs specific Samsung Galaxy smartphones to work with. This need constricts the users of Samsung Gear VR headsets because they have to use one of those specific Samsung phones which are Galaxy note 7, S7, S7 Edge, Note 5, S6, S6 edge, and S6 Edge Plus [12].

Oculus Rift headset is another example of the existing VR headsets. This headset can make the user totally immersed inside the virtual environment. In addition, it provides the user with precise picturing scenes and enjoyable interactive experience. It works with 2160 x 1200 resolution, and it is considered to be lighter than its first competitor HTC Vive, which is mentioned previously, since it only weights 470 grams. Oculus offers a package consisting of a single VR headset along with touch controllers for 798$ [13].

However, Oculus Rift DK2 is the oldest version of Oculus Rift. It is a VR headset that provides a stereoscopic 3D view which maintains excellent depth and scale. Oculus Rift DK2 has a 360º head tracking technology and an external camera for enhancing the Rift's positional tracking ability. In addition, its motherboard contains an ARM Cortex-M3 microcontroller that is used to utilize the ARM Thumb-2 Instruction Set Architecture, so the device provides full programmability and enables high performance [14].

Lastly, Google Cardboard is a VR headset that was released in 2014. It considered as one of the lightest VR headsets due to the light materials which are used to build it with. Those materials are cartoon, woven nylon, and rubber as the outer headset, and it

comes with two lenses. Also, it has some specifications that able the users to enter the virtual world [15].

3.1.3. 360º Cameras

360º cameras are common nowadays. They are provided with one or more lenses to take a certain degree photos and videos. There are many available types in the market.

As an example of the 360º cameras, the Bubl 360º camera utilize four lenses to capture the images, and it can capture 14 mega pixel spherical images and videos. It uses Wi-Fi to able the users to share the recorded videos and photos through social network. It also allows live-streaming directly to the user's PC or mobile. Bubl 360º camera also have the MicroSD card which save all the contents [16].

The 360fly 4K camera is also one of the common 360º cameras, and it has the ability to capture 16 mega pixel photos and videos. It has a water resistant up to 1 ATM, so it can capture photos and videos under the water. Furthermore, it allows user to live-steam directly, and it has a high resolution [17].

However, the Gear 360º camera records and capture videos and photos with 15 mega pixels. It has two lenses that are horizontal and vertical. Each lens captures 180º of the view, so the view in total will be 360º. It also has water resistant. The Gear 360º camera was invented for the Samsung VR headset and Samsung smartphone [18].

In addition, the LG 360º camera can capture both 360º and 180º images and videos. Also, it allows the users to capture those images with 13 mega pixels. Android 5.0 or later updates and IOS8 or later updates devices holders can use this camera. It is friendly to use with YouTube and Google Street view applications which allow the users to explore some of the 360º videos and photos and share them as well [19].

Finally, Ricoh Theta S camera is a small 360º camera that has two lenses each with 180º field of view. It can take 360º images and video, and it also support the live-streaming mode. It can be connected to the PC through a USB port or a Wi-Fi connection [20].

3.1.4. Software

There are many existing software that provide creating virtual environments like Unity3D, Open Wonderland, and Unreal.

First, Open Wonderland is considered as a 3D toolkit which can build virtual reality environments. It is a Java based application that has many services. Also, it supports many platforms such as IOS, Android, Mac, and Windows [21].

Second, Unreal is a software that support C# and Blueprints programming languages. Blueprints Visual Scripting System helps those users who are unused to deal with programming languages to create their virtual environments easily in this software. Moreover, it has many services including multiplayer capability, more detailed graphics, and flexible plugin architecture [22].

Lastly, Unity 3D software is a software that helps the developer to create the virtual scenes. It supports three programming languages which are Boo, C#, and JavaScript. It also has existed tutorials and online chat, where all of the users can communicate and help each other, on their official website. Furthermore, it has a free version which helps with lowering the cost of building a virtual application [23].

3.1.5. Existing Application

Virtual Reality is used in different fields, so many VR application were created in order to so some tasks in each of those fields.

One of these VR applications is the "Interactive Pedestrian Environment Simulator". This Simulator is used for cognitive monitoring and evaluation. It demonstrates a virtual traffic environment that helps in ensuring the safety of pedestrian and elderly people at the road. The application was built by using Unity3D software [24] and was demonstrated using Oculus Rift DK2 connected to Leap Motion and Myo by Thalmic Labs [25].

Moreover, Titan of Spare application, which is considered to be as an educational simulator, which allows students to orbit the solar system and discover the planets virtually. It provides the chance for them to learn and discuss about the space in an exciting and motivating technique. Students are allowed to perform several errands including zooming in and out the system, discovering the planets in the solar system, and gaining information about each planet. The application was developed for Google cardboard, Oculus Rift, HTC Vive, and Samsung Gear platforms [26].

Another simulator is Virtual Electrical Manual 'VEMA' which is a virtual electronics lab that facilitate the way of learning and understanding the electrical circuits. It allows the students performing circuits virtually to prevent dangerous incidents with electricity. It contains different menus such as equipment menu, capacitor menu, construction of Direct Current (DC) circuits menu, Alternating Current (AC) menu, and resonance menu. Several information is demonstrated in each menu to unsure the students' understanding. The developers of this application used Wirefusion software to build up the 3D objects in the virtual lab, and they used JavaScript as a programming language in order to interact with the 3D objects [27].

Additionally, the Chemistry Lab Application was developed as an educational animated application to provide the chance for students to perform chemical experiments virtually. It reduces the chances of having real-life injuries and causing serious damages due to the chemicals use. Developers used Autodesk Maya software for creating the 3D objects in this virtual lab. HTC Vive headset and its controllers were used to enter user to the virtual world of this application and to interact with the 3D objects inside the labs there. Unity3D software was used to build this application [28].

4. Methods

The proposed system of the application consists of both hardware and software modules. After comparing the common existing devices together and doing some calculations to see which fits our needs, the hardware module was designed. First, the common VR headsets were compared deepening on the frequency, price, angle of view, weight, resolution, connection, SDKs, platforms, and integration. Meanwhile, the common 360º cameras were compared depending on price, SDKs, number of lenses, connection, resolution, RAM, Angle, battery life, frequency, and integration. For the gesture controllers, there were few controllers on the market, so it was easy to take a decision. The most important criteria that was taken into consideration while choosing the motion controller was the accuracy because it will affect the whole project. Without an accurate controller, the application will not work in an effective way. Moreover, the tracking area of the motion sensor should be wide as much as possible to make sure that the motions will be within the range. In addition, the controller must be easy to use, so the user feels comfortable while moving. Since there is a limited budget, so the price of the motion sensor must be affordable. As a result, for the hardware implementation, Oculus Rift as VR headset, Leap Motion as gesture motion controller sensor, and Ricoh Theta S as 360º camera were chosen to be integrated together with a VR ready PC. Figure 1 shows the connections between all the devices that are used in this project. As shown in Figure 1, Oculus Rift should be connected to the PC using USB and HDMI ports. Moreover, Leap Motion controller should be connected to the PC using USB port, and Ricoh Theta S 360º camera can be connected to the PC using USB port or Theta's Wi-Fi connection that it provides. All the devices should be initialized on the PC at the beginning. The software and the needed SDKs of each device should be installed on the PC.

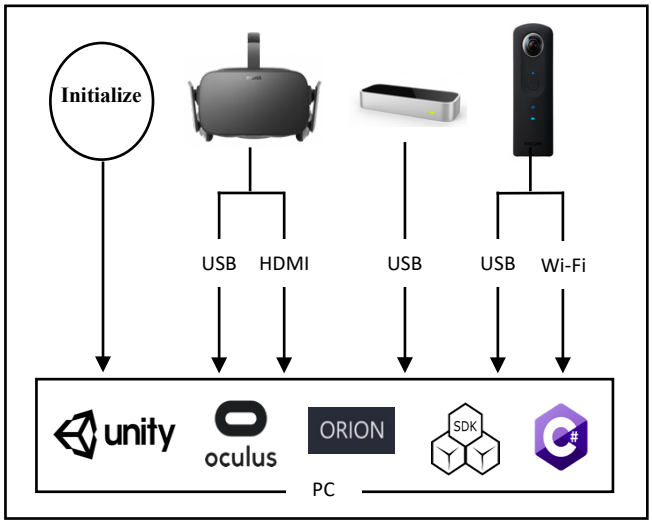


Figure 1: High Level Design Flowchart

Before choosing a software to build the application with, the specifications of the most common software that provide building a virtual environment were compared together in order to get the final decision. Those common software are Unity3D, Open Wonderland, and Unreal. Some of the main criteria of choosing the software were, the software should be easy to be dealt with and there should be existing tutorials to help with learning how to use it well. In addition, the programming languages, graphics capability, and the services that each software provides were taken into consideration. As a result, Unity3D game engine was chosen because it satisfies all the needs of this project.

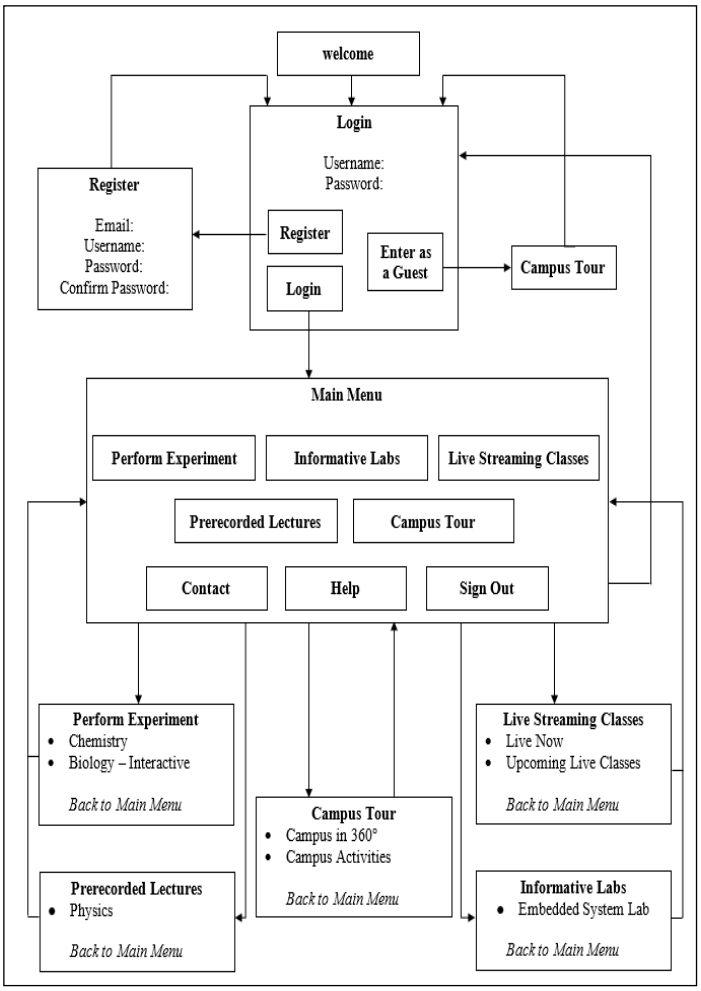


Figure 2: User Interface Hierarchy Flowchart

For the software implementation, Figure 2 shows the user interface hierarchy, and it describes the flow of the application in general. Each box in the flowchart shown in Figure 2 represents a main scene. Meanwhile, each main scene has some sub-scenes that execute a specific task such as performing experiments, watching pre-recorded lectures, having a campus tour, attending online live-streaming classes, and identifying elements and components in the informative labs. To perform those tasks, many taught courses such as Biology, Chemistry, and Physics were referred to in those scenes.

As a start, a welcome scene will be showed to the user once he plays the application. Then, he will be directed to the "Main Menu" where three main buttons which are "Enter as a Guest", "Register", and "Login" can be found. The "Enter as a Guest" button will give the user an access to have a tour around the campus only, so the user will have to register in order to get a full

access to the application. However, the application needed to be secured since it is used for an educational purpose, so a registration and a login feature were added. The user is required to fill in the needed information which are the "Username", "Email", "Password", and "Confirm Password" in the registration scene. Once the user is done with filling the information there, the system will check if the "Username" is valid or not and if the "Password" and "Confirm Password" fields are identical or not. If there is any error, the system will show a message regarding the error. Otherwise, the system will complete the registration process and add the user to the database. After that, the user will be able to login, and once he enters his username and password, the system will check the database if this user has an authorization or not. If the username is not found or the password is not correct, the system will show a message about this error. If the login is successful, the user will be directed to the "Main Menu" where he can choose to enter to one of the provided sub-scenes. If the user chose "Perform experiment" scene, he will be directed to that scene where he can choose between entering the Biology or the Chemistry labs to either interact with the objects or perform experiments virtually ending up with small quizzes to test his understanding. On the other hand, if the user chose to enter the "Informative labs" scene, he will be directed to a scene where he can visit an embedded system lab virtually to notify and learn about each equipment and their location inside the lab. For the live-streaming sub-scene, the user can enter "Upcoming Live Classes" scene to check the time schedule of the upcoming live classes. Then, if there is an available live class at any time, the user can enter "LIVE NOW" scene to attend the given lecture virtually. Moreover, by entering the "Prerecorded Lecture" scene, the user will be able to choose the "Physics Lecture" scene where a physical experiment was recorded to be displayed to the user, so he can refer to it whenever needed. In the "Campus Tour" scene, the users can choose to attend a scene where they can watch 360^0 pictures and videos of the campus facilities or visit a scene where they can display photos and videos of the campus activities taken with Ricoh Theta S 360^0 camera. Each sub-scene has a "Back to Main Menu" button, so the user can easily shift between scenes. Finally, the users can find the contact information of the university, manage the settings, get some help, and sign out from their accounts by clicking on the required button showed on the "Main Menu". Some Assets, codes, equations, object, and animations were used to implement the application.

4.1. Assets and codes

To be able to build the application using Unity3D game engine many assets and codes, which were written with different programming languages, were used. To show the hands, interact with the objects, and trigger the items inside the scenes Leap Motion assets [29] were used. In addition, Curved Keyboard [30] asset was used to show the keyboard that helps the users with entering characters in the "Login" and "Register" scenes. Furthermore, Curved UI assets [31] were downloaded into unity3D software to provide a curved interface that has 180° of view, so the interface will be flexible and easy to use. It is also a scalable interface, and it can support different VR Headsets including Oculus Rift, Gear VR, HTC Vive, and Google Cardboard. In addition, it supports many controllers other than the Leap Motion, which is the used gesture controller sensor for this project, such as the mouse, Vive controller and Oculus Touch. Lastly, Theta Wi-Fi Streaming asset [32] were downloaded to be able to have online live-streaming classes.

For codes and programing, Unity3D software support C# and JavaScript as programming languages. Both of these languages were used to write some codes in order to perform some actions inside the application. In this application, many functions were programmed to transfer the user from scene to scene, show and hide User Interface (UI) elements, display two-dimensional (2D) and 360° videos, generate animations by using Unity3D particle systems, login, register, and calculate the grades of the Biology and Chemistry experiments based on the user performance. In addition, a virtual tutor was programmed to give the user immediate instructions and feedbacks depend on his actions, so the user will be helped to move easily through the application in general.

4.2. Equations

Since there are quizzes and graded experiments in the application, few equations were used to calculate the score of the students' performance. First, the Chemistry lab has two multi-leveled experiments and a quiz. The total score of the Chemistry lab is divided to cover all those multi-leveled experiments and the final grade of the quiz. In total of a 100 scores, the first multi-leveled experiment worth 25%, the second multi-leveled experiment worth also 25%, and the quiz worth 50%. Equation 1 is used to calculate the grade of each level of the experiments. The symbol *TIME* represents the time consumption of each student to finish each experiment. Depending on the code, if the time consumption of the user is 60 seconds or below, *TIME* will get the value 60. In addition, the symbol *TRIALS* represents the number of trials the student use to complete the experiment. However, both of *TIME* and *TRIALS* mentioned in "(1)" have 50% of each level's grade which is already worth 25% of the total score as previously mentioned.

$$Score = ((50/TRIALS) + ((50*60) /TIME))/2 \qquad (1)$$

The quiz of the Chemistry lab is divided into five questions, so each one worth 10 scores out of 50% of the total score. Equation 2 is used to calculate the quiz grades. The symbol x represents the answer of each question, so if the right answer is chosen the value of x will be 1. Otherwise, the value of x will be 0. The symbol *TIME* in "(2)" represents the time consumption of the user to solve each question in the quiz. Depending on the code, if the time consumption of the user is 15 seconds or less, *TIME* will be valued as 15.

$$Score = (((15 * x) /TIME) * 10) \qquad (2)$$

For the Biology lab, it was built to be more interactive than being experimental, so the student will be only evaluated depending on the quiz grades. Equation 2, which is used to

calculate the Chemistry lab quiz, is also used to calculate the grade of the Biology lab quiz too. As the Chemistry lab, the Biology lab is divided into five questions, and each of them worth 10 scores out of 50. Since there is no experimental part on the Biology lab, the total score of it is 50.

4.3. Three-Dimensional (3D) Objects and Animations

Inside the virtual environment of this application some animations and three-dimensional objects were used. For example, Chemistry and Biology labs are animated labs which are based on animations, 3D objects, and 3D characters. In addition, the pre-recorded class in this application was built to have a combination between those 3D objects and a 2D video. Furthermore, the virtual tutor was created using 3D objects and characters. Some of those three-dimensional objects were built using tools from Unity3D game engine, and some of them were predefined from the software itself. The location, position, size, and color of those objects and animations were modified and customized to fit the project's needs. Some of these objects are interactive, so the user can hold, release, and move them inside the scenes. Those objects were created and designed in a way that appears close to the real-life objects. Hence, the objects which some of them are the labs equipment will be recognizable and observable to the user while doing the experiments or moving through the application. Moreover, fire, bubbles, and liquid drops are examples of the animations that are used in the Chemistry lab scene. Those specific animations were created using particle systems offered by Unity3D software.

5. Final Results and Achievements

As a result, Oculus Rift as a virtual reality headset, Leap Motion as a motion sensor and gesture controller, and Ricoh Theta S 360º camera were integrated together with a PC that has Unity3D game engine to build a fully interactive educational VR application. Users can interact with the application using their free hands as shown in Figure 3. The interaction between the hands and the application was done due to the detection of the users' gestures using Leap Motion controller.



Figure 3: Hardware Integration

5.1. User Interface

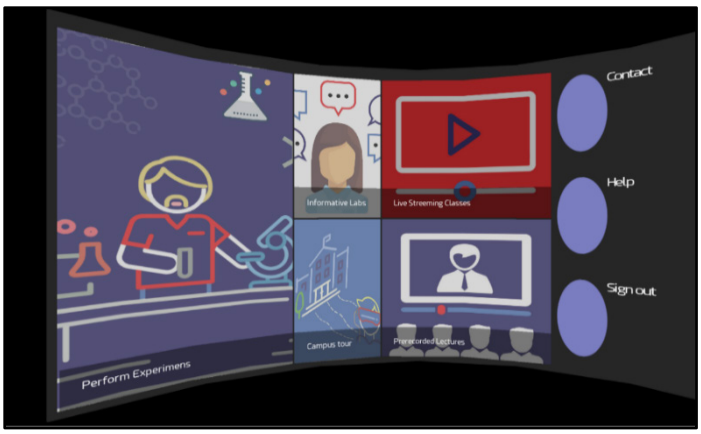


Figure 4: Main Menu User Interface

The user interface was constructed to be curved using the CurvedUI asset [31], and it was customized to fit the application's needs. Having a curved user interface can give the user 180º cylindrical view, which will make the menu fully interactive from any angle. Figure 4 shows the use of the CurvedUI asset on the "Main Menu" inside the application.

5.2. Chemistry Lab Scene

The Chemistry lab scene is divided into five levels scenes other than the main scene where the student will be able to identify the equipment that will be used during the coming levels virtually. This feature will help in solving the problem of lack of equipment and the problem of affording expensive ones in real life. In level 1 and 2, experiment 1 will be demonstrated. However, in level 3 and 4, experiment 2 will be demonstrated. Students will not be able to choose the number of the level, so he will start by level 1, pass through the next levels ascendingly, and end up with the quiz.

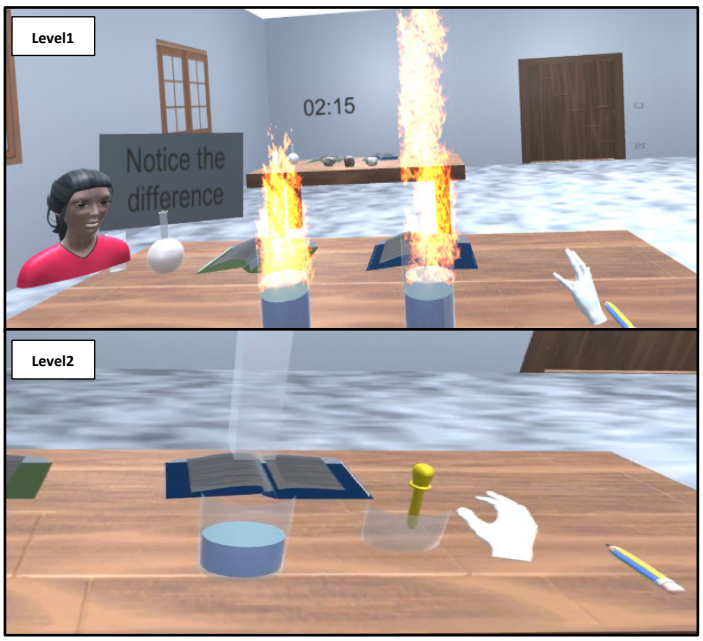


Figure 5: Chemistry Lab Scene Level 1 and Level 2

In the first level, the students will be allowed to perform a dangerous real-life experiment which is "The Reaction of Sodium with Water" that produces fire [33], so the user will try to place two virtual different sized pieces of Sodium into water and notice the difference between the reactions them. In fact, the larger the amount of Sodium the more fire will be produced as shown in Figure 5. Then, the students will check the acidity of the product in level 2 by dropping Phenolphthalein virtually on it. After dropping Phenolphthalein, if the product color turned to pink color, the product will be considered as a base.

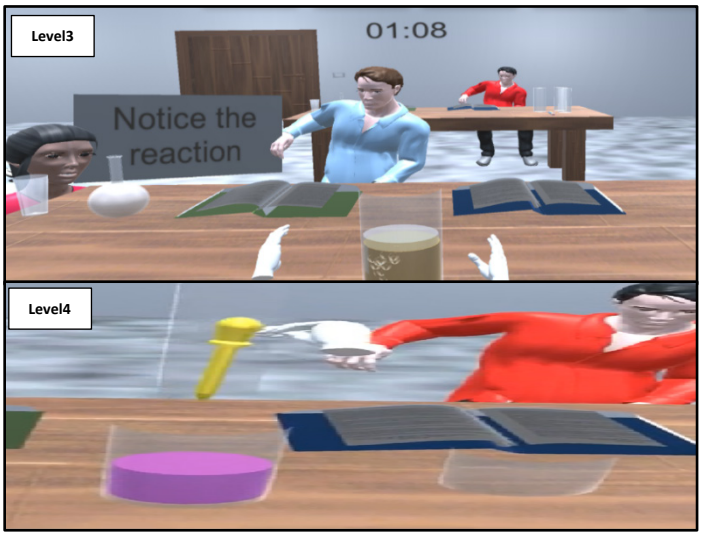


Figure 6: Chemistry Lab Scene Level 3 and Level 4

In level 3, the students will be able to recognize the reaction between Sodium and Alcohol. As level 1, the student will try to place two different sizes of Sodium inside the Alcohol. Then, they will try to notice the reaction. Unlike the reaction of Sodium with water, the reaction between Sodium with Alcohol will produce bubbles in real-life [34]. The more the amount of Sodium is used, the more bubbles will be produced. As level 2, the student will check the acidity of the product by dropping Phenolphthalein into it, and if the color of the product is changed into pink as shown in Figure 6, it will be considered as a base [35].

After the experimental part, a quiz will be provided in level 5 including five questions to test students' understanding as shown in Figure 7.

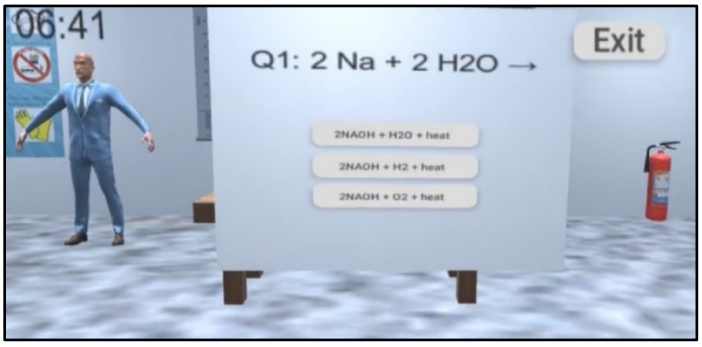


Figure 7: Chemistry Lab Quiz

Finally, the scores of experiment 1, experiment 2, and the quiz will be displayed on the score scene as shown in Figure 8.

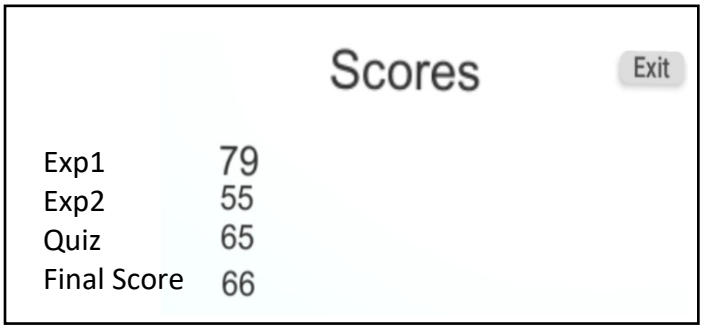


Figure 8: Chemistry Lab Scores

Regarding the applicability of VR teaching within the Chemistry lab scene, some situations are programmed to be aligned with the real-life incidents. For example, if the sodium was dropped by mistake on the ground in this application, the student can sit and pick it up as if it is real. This will give the students the chance to try and learn about how to do it in reality. Also, as mentioned previously, the different amounts of produced fire and bubbles due to the size of Sodium are examples of the applicability of VR teaching because this virtual experimental practice will lead the student to get main idea. However, some other incidents are not programmed yet. In brief, applying this experiment virtually will give the students the chance to practice the experiment and see its reaction, so if this experiment is performed in real-life the students will be more cautious. As a result of using this education supplementary virtual reality application, serious injuries will be prevented or reduced.

5.3. Biology Lab Scene

The Biology lab was built in order to help the students with learning and interacting with the biological organs virtually because it is hard to deal with them in real-life. Also, it is difficult to afford real biological organs, so this lab helps solving this problem. In the Biology interactive and informative lab scene, the user can interact with the different virtual biological objects such as human muscle, human heart, human skeleton, and a plant cell. A box of information about each object will appear to the user by triggering the objects.

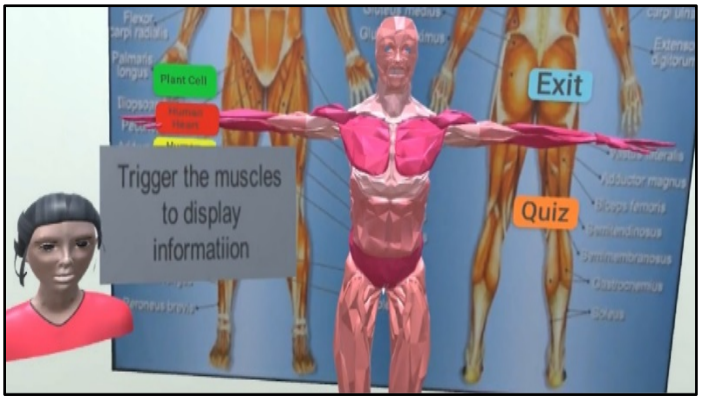


Figure 9: Human Muscles Scene

Figure 9 represents the human muscles scene, where students can see a 3D virtual human muscle module. By following the virtual tutor guidance, the student will be asked to trigger the human muscle module so that a list of information about the human muscles will be displayed in the information box.

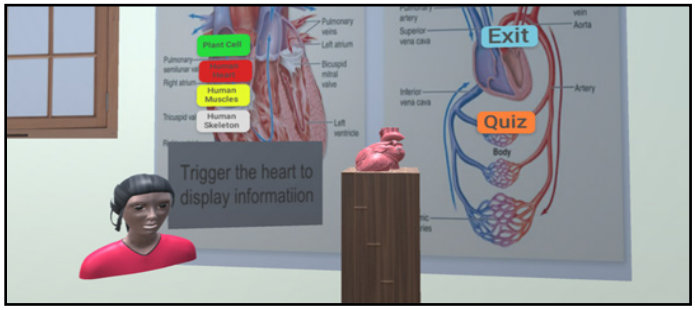
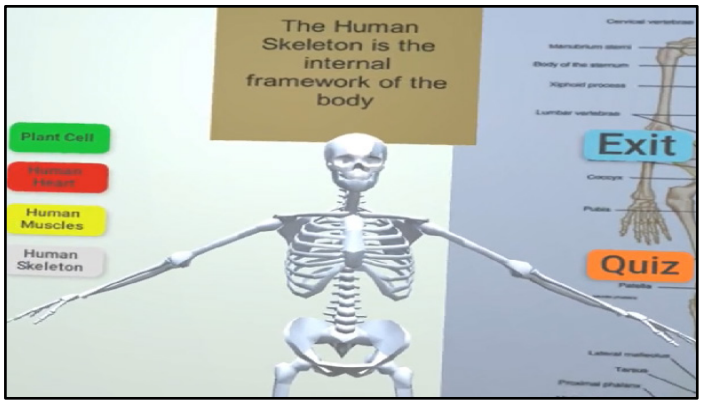


Figure 10: Human Heart Scene

The human heart scene shown in Figure 10 includes a 3D human heart module. By triggering this heart module, a list of general information about the human heart will be appeared to the student. Students will be asked in the quiz later questions which are related to the information.



:Figure 11 Human Skeleton Scene

The human skeleton shown in Figure 11 will be displayed in the human skeleton scene in the Biology lab. In this scene, students will be asked to trigger the skeleton itself, so a box of information about this body part will be shown to them.

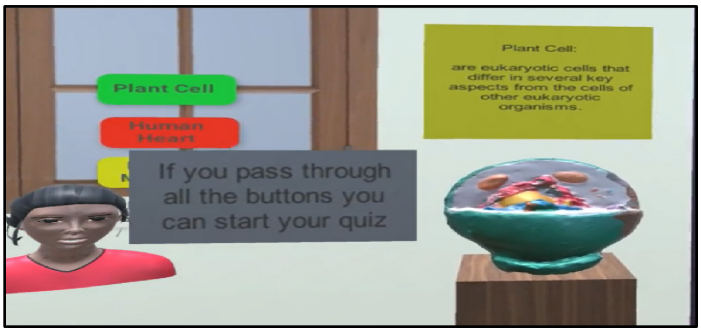


Figure 12: Plant Cell Scene

Moreover, the plant cell scene shown in Figure 12 includes a 3D plant cell module. Students can see the small components of the cell in a clear and interesting way that will help them to understand easily. Also, they can gain some information about the plant cell after looking at the information box that will be displayed when the plant cell is triggered.

After recognizing all the elements in the Biology lab, students can take a provided quiz that consist of five questions to test their understating. Then, the scores scene, where only the quiz grades will be appeared since there is no experimental part in the Biology virtual lab, will be displayed.

5.4. Pre-recorded Physics Lecture Scene

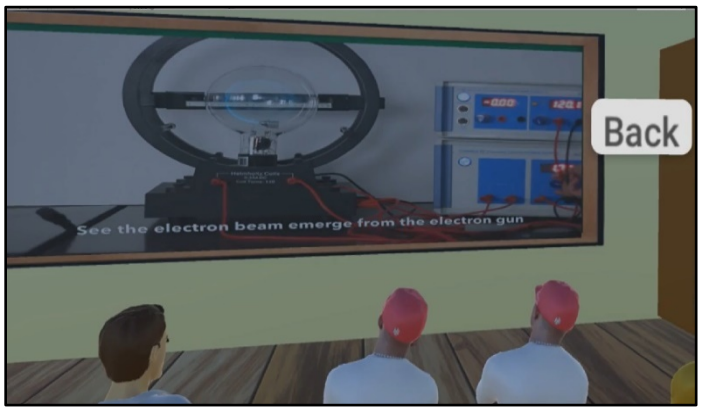


Figure 13: Prerecorded Physics Lecture Scene

A pre-recorded Physics lecture scene was created in this application to include one of the Physics experiments that need expensive equipment. Some educational societies cannot afford the equipment, so this feature provides a way to visualize what are being taught in the class with a lower cost. The recorded experiment in this scene is known as “Charge to Mass Ratio experiment” [36]. As shown in Figure 13, this scene is a combination between animated 3D characters and a 2D video.

5.5. Informative Lab Scene



Figure 14: Informative Embedded System Lab Scene

In the informative lab scene as shown in Figure 14, the user can attend a real-pictured 360º embedded system lab. This picture was taken using Ricoh Theta S 360º camera. Some floating buttons were added to the scene, so when the user press those buttons, a clear image of different equipment that are used inside the lab with its description will be shown. This lab will help students to gain more information about the use of those equipment in real-life.

5.6. Live-streaming Scene



Figure 15: Live-streaming Scene

In the live-streaming scene shown in Figure 15, the live-streaming feature was developed to provide the chance to attend virtually 360º online live classes after checking the lectures' timing on the provided schedule. To be able to have the live-streaming feature, the PC should be connected with Ricoh Theta S 360º camera's wi-fi that it provides. Having this attribute will help in increasing the number of the available seats in some particular classes that are full seated in real-life.

5.7. Campus Tour Scenes

Campus Tour considered as multiple scenes where the user can have a virtual tour inside the campus. He can see a gallery where 360º images and videos of the campus facilities are provided including the activities that have been done in the campus. All of the 360º photos and videos were taken using Ricoh Theta S 360º camera to make the user's tour be close to the reality. One of the advantages of this tour is having a virtual orientation around the campus, which will help in recognizing the campus before applying to this university or this school for example. Another advantage is, the virtual campus tour will help new students in recognizing the campus's facilities. In addition, it will allow the students, who missed the activities, to attend them virtually as if they were there.

5.8. Virtual Tutor



Figure 16: Virtual Tutor

A virtual tutor, which is shown in Figure 16, was created to guide the user inside the application. It was programmed to give immediate instructions and feedbacks depending on the user's actions. For example, it gives the user the instruction on doing the experiment inside the Chemistry lab, and it tells him well done if he passed the experiment. Also, it is used to tell the user which button to click or what movement should he do in order to perform the tasks.

5.9. Testing and Evaluation

In order to evaluate the performance of the application and to achieve the design specifications, different methods were applied to the hardware and software. First method, we moved through the application step by step and checked whether all the scenes are operating correctly or not. At the same time, the functionality of the hardware parts such as sensors were checked to make sure they are working and tracking well. In addition, all the bugs and errors that were found have been fixed.

Another followed method was, evaluating the application by multiple users and to get their feedbacks regarding their experience and the issues they faced. This includes their technical experience about the functionality of the buttons, effects, clicks, animations, and the response time of the actions. The application was tested with an approximate total of more than 50 students and 20 instructors from our university which is the American University of the Middle East, 16 instructors and 12 students from the BioSMART conference which is the Second International Conference on Bio-engineering for Smart Technologies held in Paris, and over 200 people with different positions in Knowledge Summit held in Dubai. Most of the feedbacks supported the idea of having a fully interactive educational application for practical and informative learning. On the other hand, some of them concerned about the difficulty of dealing with the virtual environment using Leap Motion gesture controller. In fact, this concern is normal because the Leap Motion controller needs some practice in order to get used to it. Some people refused the idea of replacing real teaching process with VR teaching. However, this application is an education supplementary application that will help in reducing some educational issues, so it will support the real teaching processes. At last, it was important to take these feedbacks into consideration to give the user a better experience and to improve the quality and the performance of the application. As a future plan, when there is an availability the application will be tested on people with disabilities to check their performance and to see their acceptance to the idea of this application.

5.10. New-Work

Additional to what was presented on BioSMART conference, some upgrades to the application were added. A Database using C# was created to collect data and results in one place, so it can be easily accessed and tracked. This Database keeps profile of each user whether instructor or student. Having the Database in the application will help the user to know how well he is doing, and it will help instructors in evaluating the students based on their reports easily. As shown in Figure 17 below, each user's folder contains three file which are the user's information file, the Chemistry lab scene scores file, and the Biology lab scene scores file. For the user's information file, it contains the username, email,

and password of the user as shown in Figure 17. On the other hand, the scores scenes will contain only numbers that represents the scores of each experiment and quiz in those scenes. On the other hand, the database can be used also to report the user's activity and attendance of live streaming classes and prerecorded lectures for example. However, it is a basic database, so some other upgrades will be provided to add some more professional features like those.

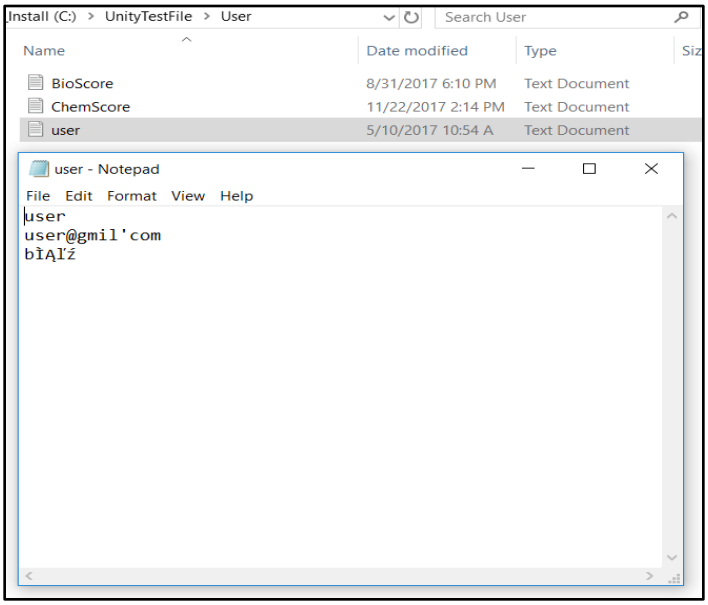


Figure 17: User's File

6. Conclusion

In conclusion, the evolution of technology succeeded in spreading the use virtual reality nowadays in several fields including education. Since that, a virtual reality application for educational purpose such as practical learning was built to solve some of the educational issues. This provided solution will facilitate the way of learning and teaching. Student's understanding and attention will be increased, and teachers will have an easier way to deliver the information to them using this application. In addition, dangerous mistakes in labs will be prevented or reduced due to the practical virtual implementation, so students will be able to practice dangerous experiments safely. Also, the number of lack of equipment problems will be decreased since those equipment will be provided simply inside the virtual environment. Nevertheless, students with disability will be involved in performing experiments with less movements struggles. Furthermore, this application is supporting gaining information in an interesting way because of the use of virtual reality technology. As an outcome, Oculus Rift as a virtual reality headset, Leap Motion as a motion sensor and gesture controller, and Ricoh Theta S as a 360° camera were integrated together with Unity3D game engine to produce this educational tool that can help in developing the educational field.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This project was supported and Funded by the American University of the Middle East (AUM) where we spent time doing the project at the Research and Innovation Center with endless support from the professors, instructors, and engineers. We would like to show our gratitude to them for assisting, helping, and shearing their pearl of wisdom with us to complete this project.

References

[1] S. AlAwadhi N. AlHabib, D. Murad, F. AlDeei, M. AlHouti, T.Beyrouthy, S. AlKork., "Virtual reality application for interactive and informative learning," 2017 2nd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, 2017, pp. 1-4. doi: 10.1109/BIOSMART.2017.8095336
[2] G. Leman and W. Tony, "Virtual worlds as a constructivist learning platform: evaluations of 3D virtual worlds on design teaching and learning," Journal of Information Technology in Construction (ITcon), pp. 578-593, 2008.
[3] P. Brey, "Virtual Reality and Computer Simulation," The handbook of information and computer ethics, pp. 361-363, 2008.
[4] G. Burdea and P. Coiffet, Virtual Reality Technology, Canada: John Wiley & Sons, 2003.
[5] G. V. B. Yuri Antonio, "Overview of Virtual Reality Technologies," Southampton, United Kingdom, 2013.
[6] G. Daniel, "Will gesture recognition technology point the way?," vol. 37, no. 10, pp. 20 - 23 , 2004.
[7] N. Rachel, "Armband adds a twitch to gesture control," *New Scientist,* vol. 217, no. 2906, p. 21, 2013.
[8] X. Zhang, Z. Ye and L. Jin, "A New Writing Experience: Finger Writing in the Air Using a Kinect Sensor," vol. 20, no. 4, pp. 85 - 93, 2013.
[9] Higgins, Jason Andrew, Benjamin E. Tunberg Rogoza, and Sharvil Shailesh Talati. "Hand-Held Controllers with Capacitive Touch Sensors for Virtual-Reality Systems." U.S. Patent Application 14/737,162, 2016.
[10] "Leap Motion Unity Overview," Leap motion, 2016. [Online]. Available: https://leapmotion.com.
[11] L. Prasuethsut, "Why HTC Vive won our first VR Headset of the Year award," WAREABLE, 2 November 2016. [Online]. Available: https://www.warea [2]ble.com.
[12] G. Carey, "Spec Showdown: Oculus Rift vs. Samsung Gear VR," 6 January 2016. [Online]. Available: http://www.digitaltrends.com.
[13] "intoducing Oculus Ready PCs," 2016. [Online]. Available: https://www.oculus.com.
[14] B. Lang, "Oculus Rift DK2 Pre-order, Release Date, Specs, and Features," Road to VR, 5 October 2016. [Online]. Available: http://www.roadtovr.com.
[15] "Cardboard," google, 2016. [Online]. Available: https://vr.google.com.
[16] "Be a part of 360 Spherical Movement," Bublcam, 2015. [Online]. Available: https://www.bublcam.com.
[17] "360fly 4K Specifications," 360fly, 2016. [Online]. Available: https://360fly.com.
[18] "GEAR 360," samsung, 2016. [Online]. Available: http://www.samsung.com.
[19] "LG 360 CAM," LG, 2016. [Online]. Available: http://www.lg.com.
[20] "High-spec model that captures all of the surprises and beauty from 360°," THETA, 2016. [Online]. Available: https://theta360.com.
[21] J. Kaplan and Y. Nicole, "Open Wonderland: An Extensible Virtual World Architecture," vol. 15, no. 5, pp. 38 - 45, 2011.
[22] G. A. Kaminka, M. M. Veloso, S. Schaffer, C. Sollitto, R. Adobbati, R. Adobbati, A. Scholer and S. Tejada, "GameBots: a flexible test bed for multiagent team research," *Communications of the ACM - Internet abuse in the workplace and Game engines in scientific research,* vol. 45, no. 1, pp. 43-45, 2002.
[23] "Unity 3D," Unity 3D, 2016. [Online]. Available: https://unity3d.com.
[24] J. Orlosky and M. Weber, "An Interactive Pedestrian Environment Simulator for Cognitive Monitoring and Evaluation," pp. 57-60 , 2015.
[25] N. Owano, "Thalmic Labs' Alpha users explore Myo with Oculus," 2014.
[26] "Titans of space," 2016. [Online]. Available: http://www.titansofspacevr.com.
[27] V. M. Travassos and M. C. Ferreira, "Virtual labs in electrical engineering education - The VEMA environment," Information Technology Based Higher Education and Training, 2014.

[28] C. Smith, "Chemistry Lab VR," 2016. [Online]. Available: https://devpost.com.
[29] "Leap Motion Unity Overview," Leap motion, 2016. [Online]. Available: https://leapmotion.com.
[30] "Unity 3D," Unity 3D, 2016. [Online]. Available: https://unity3d.com.
[31] "Curved UI - VR Ready Solution To Bend / Warp Your Canvas!," Unity 3D, 2016. [Online]. Available: https://www.unity3d.com
[32] High-spec model that captures all of the surprises and beauty from360°," THETA, 2016. [Online]. Available: https://theta360.com
[33] C. Bobbert, "Solvation and chemical reaction of sodium in water clusters," The European Physical Journal D - Atomic, Molecular, Optical and Plasma Physics, vol. 16, no. 1, p. 95–97, 2001.
[34] J. Clark, "REACTING ALCOHOLS WITH SODIUM," 2009. [Online]. Available: http://www.chemguide.co.uk/.
[35] Repp, Julia, Christine Böhm, JR Crespo López-Urrutia, Andreas Dörr, Sergey Eliseev, Sebastian George, Mikhail Goncharov, "PENTATRAP: a novel cryogenic multi-Penning-trap experiment for high-precision mass measurements on highly charged ions." Applied Physics B 107, no. 4 (2012): 983-996.
[36] Al Kork SK, Denby B, Roussel P, Chawah P, Buchman L, Adda-Decker M, Xu K, Tsalakanidou F, Kitsikidis A, Dagnino FM, Ott M, Pozzi F, Stone M, Yilmaz E, Uğurca D, Şahin C "A novel human interaction game-like application to learn, perform and evaluate modern contemporary singing: "human beat box"". In: Proceedings of the 10th international conference on computer vision theory and applications (VISAPP2015), 2015
[37] Ulmer, S., C. Smorra, A. Mooser, Kurt Franke, H. Nagahama, G. Schneider, T. Higuchi et al. "High-precision comparison of the antiproton-to-proton charge-to-mass ratio." Nature 524, no. 7564 (2015): 196, 2015.
[38] Al Kork S, Amirouche F, Abraham E, Gonzalez M. "Development of 3D Finite Element Model of Human Elbow to Study Elbow Dislocation and Instability". ASME. Summer Bioengineering Conference, ASME Summer Bioengineering Conference, Parts A and B ():625-626, 2009
[39] E. Saar, "Touching reality: Exploring how to immerse the user in a virtual reality using a touch device," p. 43, 2014.
[40] D. D. Carlos, Corneal Ablation and Contact Lens Fitting: Physical, Optical and Visual Implicatios., Spain: Universidad de Valladolid, 2009.
[41] Patrick CHAWAH et al., "An educational platform to capture, visualize and analyze rare singing", Interspeech 2014 conference, Singapore, September 2014

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 74-82 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Visualizing Affordances of Everyday Objects Using Mobile Augmented Reality to Promote Safer and More Flexible Home Environments for Infants

Miho Nishizaki*

Department of System Design, Tokyo Metropolitan University, 191-0065, Japan

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 20 October, 2018
Online: 01 November, 2018

Keywords:
Augmented Reality (AR)
Affordance
Visualization
Accident
Development
Infant

ABSTRACT

This study presents a prototype augmented reality (AR) application that helps visualize the affordances of everyday objects for infants in their home environments to prevent accidents and promote development. To detect and visualize affordances, we observed 16 infants, 4 to 12 month of age, on how they perceived and handled common objects in their homes in Tokyo, Japan and in Lisbon, Portugal. Based on the longitudinal data, we developed an AR application for handheld devices (iPhone and iPad) and tested two types of vision-based markers. Ten types of basic objects were selected from the results of the observation and embedded into the AR markers. AR contents illustrated infants' actions toward objects based on actual video data recorded for security purposes. To confirm the prototype's advantages and improvements, informal user interviews and user tests were conducted. The results demonstrated that the prototype reveals the relationship between infants and their home environments, what kinds of objects they have, how they perceive objects, and how they interact with these objects. Our study demonstrates the potential of this application's AR contents to enable adults to better understand infants' behavior towards objects by considering the affordances of everyday objects. Specifically, our app assists in improving the perspective of adults who live with infants and promotes the creation of more flexible and safer environments.

1. Introduction

1.1. The Environment and Augmented Reality

Augmented reality (AR) technology enables us to visualize our living environment in novel ways. In 1968, Ivan Sutherland created the first AR experience generated using a head-mounted display [2], [3]†. In the early 1990s, the term "augmented reality" was coined with the first technological development underlying augmented devices [4], [5]. Azuma described AR as systems with three technical components: (1) use of real and virtual elements, (2) real-time interactivity, and (3) 3D registration [6]. The third characteristic "3D registration," refers to the system's ability to anchor virtual content in the real world, that is, in a part of the physical environment [7].

AR enhances user perceptions of and interactions with the real world by seamlessly merging virtual content with reality [5], [7]. Whereas virtual reality technology replaces reality by creating a whole immersive virtual environment, AR enhances the reality of an experience without changing it by projecting partially virtual/digital content in a non-immersive way.

Current AR devices require four elements: displays, input devices, tracking, and computers. Displays include head-mounted, handheld, and spatial displays. Mobile AR systems use handheld displays such as cellular phones, smartphones, and tablets. The first mobile-phone-based AR application was demonstrated in 2004 [8], and the mobile AR market has since grown consistently [6], [9], [10–12]. Such applications became well established with the advent of Pokémon Go in July 2016. This successful location-based mobile game required users to find virtual characters in the real world, and it immediately became an international sensation. It infiltrates users' physical lives and changes their locomotor behavior with potential effects to their health. Pokémon Go was launched for the Apple Watch in December 2016 and helped further promote health effects. Even now, in 2018, this game remains popular [13].

AR application development for mobile devices has become a large and fast-growing area as AR glasses and head-mounted displays are being improved further. The AR market is valued at $83 billion, whereas the VR market is valued at only $25 billion

*Miho Nishizaki, Department of System Design, Tokyo Metropolitan University, 191-0065, Japan. †This paper is an extension of work originally presented in 2017 Intelligent Systems Conference (IntelliSys) [1].

[14]. Azuma suggested the following potential application areas for AR: medical visualization, equipment maintenance and repair, annotation, robot path planning, entertainment, and military aircraft navigation and targeting [7]. Thus far, AR has been used in fields such as architecture, clinical psychology, cognitive and motor rehabilitation, education, and entertainment. AR has become popular; nonetheless, more research is needed on how to use it to better living experiences. The use of AR could be increased by improving device and recognition technologies. However, few studies and applications of AR currently exist for solving everyday issues.

1.2. Accidents and Developments in Home Environments

Young children spend most of their time at home while they grow up. Although most people think of accidents as things that happen outside the house, accidental injuries among infants and young children at home have been a major issue worldwide. For example, 14.5% of aged 0–4 years children's deaths are caused by home accidents in Japan [15]. Similarly, in the UK and the USA, the 0–4 year age group is at the highest risk of home accidents and the accidents at home are the leading cause of deaths [16–18]. Therefore, many studies have explored how to prevent accidents at home [19–23]. However, suitable methods for collecting actual data at home and for conducting quantitative and qualitative analyses have not yet been established.

Most injury alerts are based on hospital reports [24–27]. The Japan Pediatric Society publishes "Injury alerts" since 2008 and "Follow-up articles" since 2011 based on members' reports on their website. Although such information is useful and important, it is often incomprehensible to ordinary people. Therefore, such information needs to be presented clearly and be accompanied by detailed explanations. Several studies have highlighted that accidental injuries suffered by children at home are preventable [21–24]. This study uses an AR system to visualize and simulate infants' behavior in a concise and clear manner to provide caregivers of children with insights into their own living environments.

1.3. Visualizing Possibilities-Affordances

It is too difficult to protect infants from all dangerous possibilities. At the same time, just to prevent accidents, children may be banned from exploring various things at all stages of development. It is important to avoid overly protective behavior or overcontrolling children. Therefore, we employed James Gibson's theory of affordance [28,29] to deal with infants' unexpected behaviors at home during growth. The affordance is an essential concept of ecological approach; it refers possibilities or opportunities for behavior that the environment offers an animal. Gibson proposed that an affordance is a fact in the environment as well as a fact embodied by an act. It implies the complementarity of the animal and the environment [29]. Thus, the affordances of the home environment in which we spend most of the time and our way of life are inseparable. Furthermore, according to Gibson, "the affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill [29]."

Perceiving the information of affordances is not the same even if people share the same objects or space. Although adults know the way to use objects at home as designed, infants explore every object and find a variety of possibilities. We do not have to know the name of the object or how to use it to perceive what it affords. Empirical studies on individual affordance have been advanced in a series of pioneering studies since 1960 [30–34]. However, researchers have defined affordance in a variety of ways. The generalization methods are still being investigated as they have not yet been thoroughly established [35–39]. Experiments on infants' natural activities have been conducted in a laboratory playroom [40–42]; however, these kinds of studies are few in number and little is known about everyday home environments. Therefore, the current study aims to present a method adaptable to the home environment that would help visualize resources to support healthy growth of children from a variety of aspects.

This study has three main objectives: First, to identify infants' unique behaviors in home environments by the longitudinal observations during the first year; second, to examine data for two countries whether there is a difference or not; and third, to present an AR mobile application prototype that enables users to simulate infants' behaviors toward objects that assists caregivers to understand intuitively and hence provide better environments for infants.

2. Methods

2.1. Participants

We recruited ten healthy infants (6 males, 4 females) in Japan and six healthy infants (4 males, 2 females) in Portugal from middle or upper middle-class families for the longitudinal observations. In Japan, eight Japanese families and one Hungarian family were from the Tokyo area. In Portugal, all Portuguese families were from the Lisbon area.

2.2. Procedure

Observation. Longitudinal observations were conducted to clarify how infants behave at home during their first year. We observed the activities of 16 infants at their homes from 4 to 12 months of age. An experimenter or infants' caregivers recorded the infants' behaviors with digital video cameras (EX-ZR500; Casio Corp) for over 60 min per month. The recordings needed to cover several natural behaviors (e.g., it is preferable to not film only one activity such as napping). When caregivers recorded the videos, they were asked to send the recorded video data to us every month via a mail or file transfer service. Although parts of monthly data from both countries were unavailable because of family reasons, we collected 122 h of digital videos for Japan and 53 h for Portugal in total.

Design and Development. This process consisted of four steps: (1) sketch and paper prototyping; (2) design of an iOS application using Adobe Illustrator CC and Photoshop CC; (3) development of the iOS application using OpenCV 2.4.9, Objective-C, and Swift 4.0; and (4) creation of AR movies based on videos from the observations using Adobe Photoshop CC and Adobe Lightroom CC. All AR movies were illustrated manually as line drawings to protect privacy and make it easy to zoom in infant–object interactions. More than three drawings were needed for every second and each playtime is less than 60 s. Line drawings were digitized using Lightroom and Photoshop. Every image was saved at 448 × 336 pixel resolution and combined and exported as

QuickTime movies. An example of line drawings of a cabinet with an infant (11 months, 17 days) is shown in Figure 1.

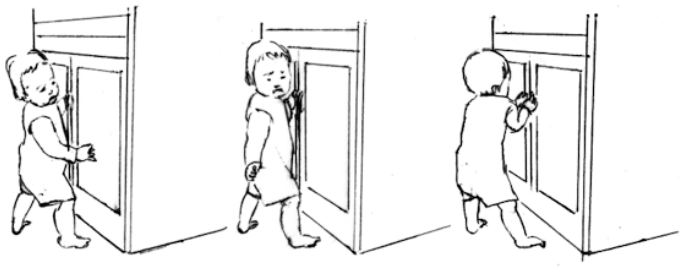


Figure1. Example of line drawings (cabinet).

2.3. Data Coding

Two coders scored the infants' attribute, house style, behaviors, postures, milestones, ages, places, durations, objects/people (if infants directly touched), and surfaces from digital video recordings by using the Datavyu coding software [43], with reference to the caregiver's comments or notes.

2.4. Ethical Considerations

All procedures for this study were approved by the relevant institutional review boards in consideration of ethical standards for research on human subjects. Informed consent was obtained from all families.

3. Results

3.1. Reality from Observation

A prototype AR application was developed to enable the simulation of children's actions in the first year based on recorded data from observations. We observed infants' spontaneous interactions between surrounding objects at their home from the recorded data. Two infants (one Japanese and one Portuguese) were excluded from the analysis because the data for these babies were not enough 9 months.

Number of Objects at Home. Infants' interactions with surrounding objects were counted using the observational data. All objects that infants chose to interact with met the following two requirements: 1) Infants directly touched the object by him/herself, and 2) objects observed at infants' homes were items that were used daily, such as furniture.

On the whole, Japanese infants, between 4 to 12 months of age, encountered 292 types of objects at their homes. As for Portuguese infants, 177 types objects were observed at their homes. Figure 2 shows the kinds of objects that exist and how many objects exist in each country. According to the ecological psychologist James Gibson's definition of the term objects, in this study, we categorized the two types of objects: detached objects and attached objects [29]. A detached object can be displaced from the surface while an attached object cannot. We identified 262 detached objects including furniture, daily used items, toys, and 30 attached objects including enclosures/corners, aperture, steps, and concave/convex objects, in our recordings in Japan. Similarly, in Portugal, we observed 153 detached objects and 24 attached objects. A number of differences were found in the detached objects, especially daily used items and toys. Japanese infants could reach more stationary items (e.g., crayons, erasers, scales, pencils, pens, drawing boards, papers, and notebooks), and kitchen items (e.g., cups, dishes), laundry items (e.g., clothespins, hangers, laundry basket), and room conditioner (e.g., thermometer, humidifier, hot water bottle, heater, stove) than Portuguese infants. Regarding toys, more handmade toys, digital video games, and character items were observed in Japanese homes than in Portuguese homes.

Figure 3 shows the percentage of the number of objects that infants interacted with during the observation period in each country. The detached objects, especially the daily-used objects, were observed most frequently among all categories in both countries. It suggested that the tendency of interaction with objects of these categories is similar even though the country is different.

Based on this result, we selected ten objects which were common, continuous, and frequently-observed, for this prototype AR application. Detached objects included cabinet, chair, cushion, futon/mattress, sofa, and table. Attached objects included bathtub, door, threshold, and wall.

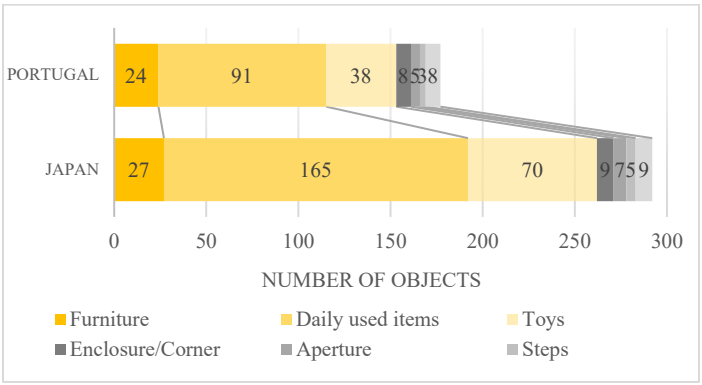


Figure 2. The number of objects that infants directly touched at home in each country. The yellow bars indicate detached objects and the monochrome bars indicate attached objects.

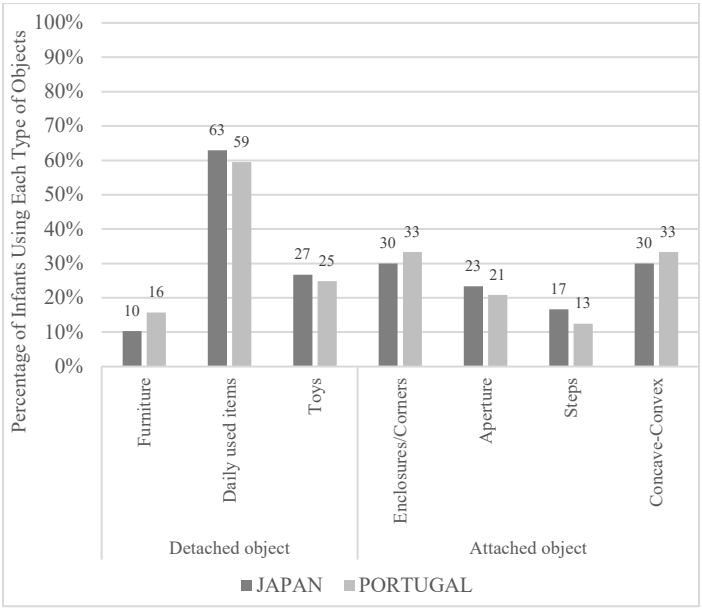


Figure 3. Percentage of each object-category in each country.

Actions and Perceptions. The results of the observations showed infants' spontaneous activities involving objects in their homes. It combines both developmental possibilities and accident risks. To clarify how often every action occurred or were supposed to occur, we counted the number of possibilities of development, accidents, and both accidents and development from all recordings. In this case, we measured the possibilities because we attempted to protect the infant from the accident before accidents occurred during observation. Therefore, the moments also applied to such cases: Parents say "watch out" to the infant or supported the infant's posture before falling.

The number of sessions were 1771 in total (M = 196.78, SD = 110.17) for Japanese infants, and 910 (M = 182, SD = 117.64) for Portuguese infants. The conditions of three types of selections are as follows. Accidents were linked to injury-inducing behavior such as losing balance (slip, trip, and fall down), drinking or eating harmful substances/objects, and touching something dangerous for infants. Demonstrations of preventative actions by people in the immediate vicinity before the occurrence of accidents are also included. This includes the same aforementioned actions. Development is associated with motor, perceptual, and cognitive progress to a greater degree than before. Both accident and development are recognized in the same session.

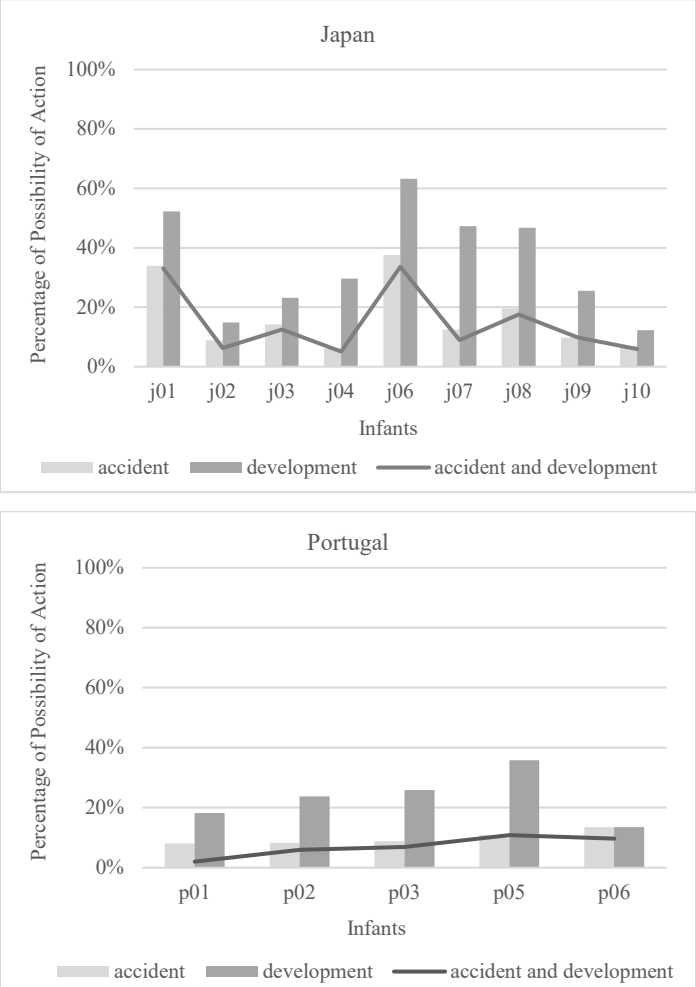


Figure 4. Percentage of action possibilities related to development, accidents, and both in each country per infant.

In both countries, for all sessions, the number of developments was higher than accidents and the combination of accident and development. Figure 4 shows the percentage of action for possibilities of development, accidents, and both in every infant's total sessions per country. The values for accidents and both accidents and development were close in most cases. It indicates that the accidents tend to involve developmental issues that might have affected the situation.

There were over 30% of possibilities of accidents for two infants (Japan: j01 and j06). Common features in these incidents included the detached objects, furniture, and daily used items in both countries. These objects were adapters, cords of light/AV equipment/laptop PC/phone, pencils on tables, cream of tube package, washing machines, bathtubs, sinks, fence, tripods, strings, steps, low tables, cabinets, and doors. However, the only observed baby products that caused accidents were the walker and the high chair. Needless to say, the accidents are related to motor developmental changes; however, the results suggest that many normal products that are designed for a particular purpose were perceived differently by infants, and this difference may cause accidents (e.g., dressers, cabinets and washing machines sometimes offer opportunities to touch, hang on, climb on, and enter inside, for infants. However, there are a number of accidents related to washing machines in particular [44]).

3.2. AR application and vision-based tracking

Vision-based tracking is defined as registration/recognition and tracking approaches [5, 6]. In this study, two types of vision-based marker codes were examined (Figure 5). In Figure 5 (a), OpenCV, MATLAB, and Illustrator, are used for compounding dots and object icons. Figure 5 (b) uses Vuforia and Illustrator for combining textures and icons. Both markers can be attached to real-world objects that can be printed using a printer.

We measured the tracking times to determine if there is a difference between two types of markers (Figure 5) along the two directions of surface (horizontal and vertical) under two lighting conditions (A: Fluorescent light, B: Flat-surface Fluorescent light.). Tables 1 and 2 show the relation between velocity and illuminance of the two directions of surface and the distance from marker to device. To confirm the accuracy of the marker recognition related to two differnt light condition, we used an illuminance meter (KONICA MINOLTA T-10). Differences between two markers at distances greater than 30 cm, regardless of light conditions, were observed (Table 1 and Table 2). Marker (b) indicates more unstable properties than marker (a) in both conditions. Moreover, additional differences were found between the two orientations of the surface under the lighting (B). Although the illuminance was sufficient to capture an AR maker, the tracking time was slightly longer for vertical surface than for horizontal surface. This suggests that surface orientation affects tracking in the same manner as other AR marker systems.



Figure 5. Examples of AR markers. Left: dot-type marker (*a),* Right: texture-type marker (b).

Table 1. Change of velocity related to illuminance and distance between marker and device under fluorescent light (A).

Type of Marker	Surface	lx	20cm	30cm	40cm
(a)	vertical	289	3 sec	4-5 sec	3 sec
	horizontal	330	3 sec	4-5 sec	3 sec
(b)	vertical	290	5 sec	—	—
	horizontal	300	5 sec	—	—

Note. Fluorescent light's product # is FPL36EX-N, 4 tubes × 20, 2900 lm/tube.

Table 2. Change of velocity related to illuminance and distance between marker and device under flat-surface fluorescent light (B).

Type of Marker	Surface	lx	20cm	30cm	40cm
(a)	vertical	916	3-7 sec	4 sec	3 sec
	horizontal	865	2 sec	2 sec	3 sec
	vertical	427	4-5 sec	4 sec	4-6 sec
	horizontal	330	1-2 sec	1-2 sec	1-2 sec
	vertical	57.5	3 sec	3-5 sec	3 sec
	horizontal	53	2-3 sec	2-3 sec	3 sec
(b)	vertical	916	8 sec	—	—
	horizontal	865	5 sec	—	—
	vertical	427	3 sec	—	—
	horizontal	330	3 sec	—	—
	vertical	57.5	5 sec	—	—
	horizontal	53	3 sec	—	—

Note. Flat-surface fluorescent light's product # is ELF-554P, 4 tubes × 1, F55bx/Studiobiax32, 4100 lm/tube, 3,200 K.

Overview. Figure 6 depicts an overview of how to use the prototype application. First, in order to watch AR movies, the user needs to put markers on objects at his/her home. After starting the application, the user can tap the camera icon. Once tapped, the camera window appears and the user capture the AR marker in a manner similar to taking a photograph. When the camera recognizes the marker, a pop-up showing the list of movies according to the age in months is shown. As the user select the movie, AR movie begins to play.

Every marker contains approximately 10 AR movies that illustrate the interactions of infants between 4 and 12 months of age, with objects. The prototype enables the user to simulate how infants interact with the environment and know the difference according to the objects and infants' age. Furthermore, users can also learn about infants and objects. We introduced a notification window of an example for positive and negative tips in Figure 7. Users can find further information as they tap the object icon. In the case of risk information, it is composed with recorded conversations from the data, comments from caregivers, and links to websites.

Figure 8 shows another example of attached objects for further references. This small dictionary was implemented in the application to serve as an introduction for the application users. The users were able to select the content to understand objects at home from the perspective of infants from the floor plan, attached objects, and detached objects.

3.3. User Experience

To determine whether our application is functional, informal user interviews and tests were conducted with 142 people from six countries (Italy, Japan, Portugal, Spain, UK, and USA) from 2013 to 2017 as part of testing phase for the prototype. The application has been improving since 2013 with the addition of longitudinal observation data and updating functions. The interviewees were spread across 18 fields and nine occupations, and consist of architects, designers, developers, engineers, office workers, researchers, students, teachers, and infants' parents. The interviews revealed the advantages of the prototype and possible improvements. To examine what potential users need to know about infants, we conducted a SWOT analysis based on all results of the interviews. Table 3 shows a summary of this analysis.

As a result of the interviews, the application's advantages were categorized into strengths and opportunities. These indicate that visualization of infant behavior using mobile devices using AR technology was able to gain adults' understanding of infants' and the home environment's specific affordances as a whole. Regardless of nationality, occupation, and sex, the interviewees showed interest in the infants' behaviors as "unexpected" interactions with daily used items.

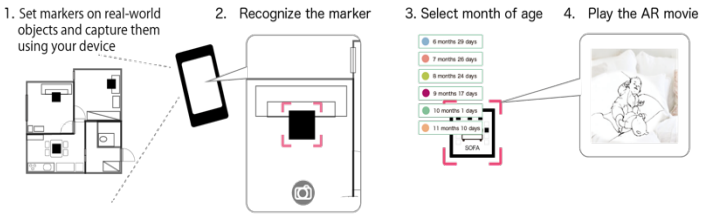


Figure 6. How to use the application.

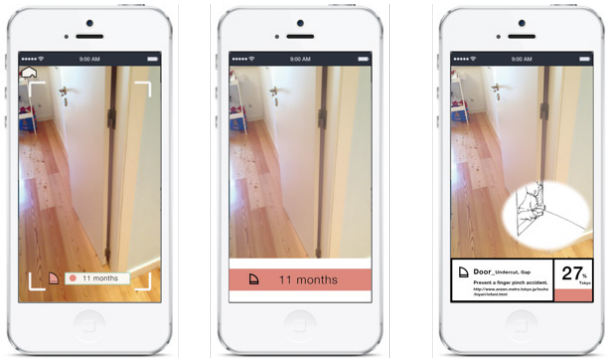


Figure 7. Example of notifications.

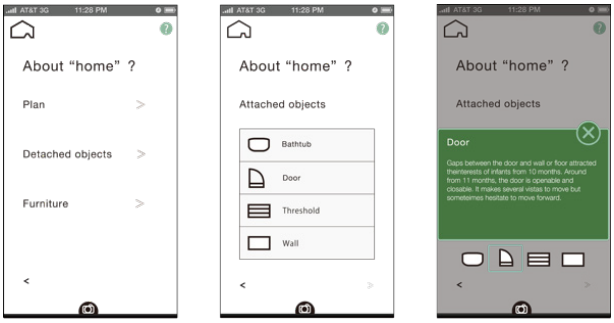


Figure 8. Example of content with pop-up windows of an attached object.

Table 3. Summary of SWOT analysis for the prototype.

CAT	Strengths	Weaknesses	CAT
【Usability】			
R, O, P, S	Instant access to information and camera.	Improve recognition.	R
R, P	Protected privacy with line drawings.		
P	Add more function (e.g., recommendation of products).		
【Method】			
R, P, D	All animations are based on longitudinal data is great.	Hard to collect longitudinal data and archive data.	R
		Rich video data; however, difficult to analyze.	R
		Difficult to draw all movies.	R, D
		Need collaborators from several fields.	R
【Effect】			
R, P, S	Easy to understand the characteristics of infants.		
R, P	Reduced risk of inaccurate perceptions or misperceptions.		
R, E, P	Increase interests in infants' behaviors for adult.		
R, P, D	Adds opportunities to reevaluate development.		
R, D	Interactive applications. Enhance communication.		

CAT	Opportunities	Threats	CAT
【Method】			
R	Aid in elderly people's accident analysis.	Radical change in technology.	R, E
R	Applicable to use for autistic children.	Frequent update of programming languages and libraries.	E
R, E	Increase adaptation to wearable devices.		
【Effect】			
R, P	Increase safety to reduce preventable accidents.		
R, D	Promote making the home childproofing.		
O, P	Good for adults, especially new parents.		
R	Potential in the psychology/healthcare/educational fields.		
【Possibility】			
R, D, E	Can bridge several fields through collaboration.	Big data of hospital or healthcare areas.	R
R, P	Collaborative products with this application.		

Note. Interviewees' categories (CAT) are as follows: D = designers/architects; E = engineers/developers; O = office workers; P = parents; R = researchers; S = students.

In terms of strengths, almost all interviewees supported the method of converting digital recorded videos to line drawings. In particular, line drawings ensured a greater feeling of security for the infants' parents compared to embedded raw videos. In terms of reality and security, this method is assumed to be appropriate at the present time. With respect to usability, we found that all interviewees were familiar with the operation of smartphone devices. This suggests that the development of applications for handheld devices are appropriate for everyday use. In the case of interface, the application's simple graphic and motion design enabled users to access the desired information and to reach the camera icons easily. However, it was also observed that tracking times were affected by the lighting conditions, especially in the indirect lighting condition, it appeared to unstable at times.

Regarding opportunities, researchers who specialized in robotics, design, architecture, and computer science suggested possibilities for expanding the scope of the application to several other fields, such as preventing accidents among the elderly or children with autism. Furthermore, derivatives of the product possibilities based on the application and the longitudinal data were suggested.

Improvements are suggested in the sections where weakness and threats are discussed. Most of weaknesses in the collection of data and its processing were identified. Researchers concerned that data collecting in natural settings are rich resources for visualization; however, they also referred to the difficulty in analysis compared to the data from laboratory experiments. In contrast, this study is considered to be better at collaborating with other fields because of its interdisciplinary work; hence, engineers suggested technical improvements of the AR technology through continuous updates and researchers pointed out the analysis methods.

4. Discussion

The first year of a child's life is full of indefinite positive and negative opportunities that enable children to perceive, act, and learn. These affordances can be changed based on the relationship between an infant and the surrounding environmental conditions, namely, age, sex, motor skills, body scale, order of birth, object, surface, layout, space, and other people. The affordances of the environment that are available to infants are different from those available to adults and infants' actions are also unpredictable to adults. This may result in carelessness or overprotection. Our AR application attempts to consider both positive and negative affordances. In this respect, our application differs from previous injury prevention systems. This feature of the prototype could make a contribution by changing "preventable" incidents into prevention and even more promotion, whichever is appropriate.

4.1. Everyday Life and Objects from Observational Data

In this study, we examined the data from two countries based on longitudinal observations conducted in Japan and Portugal during the first year and measured infants' activities and objects' types and numbers that they directly encountered. Even though the number of infants were smaller than those in our cross-sectional method, the 175 hours of recorded data captured infants' changing processes in the real world. Therefore, we found the common tendencies in the daily use of products that tend to be associated with the accidents. This means that the infants' interactions with these objects are sometimes unpredictable to adults and vary depending on the conditions of the body-environment relationship, layout of objects and people. Adopting only an adult viewpoint in this aspect leads to carelessness or overprotection.

Concretely, this prototype attempted to visualize how infants cope with ten types of normal objects based on the recorded data as evidence. This visualization was enabled using AR technology and facilitated adults' intuitive understanding by focusing on the decisive moments of the infants' action toward the object.

Essentially, this application provides objects' potential issues and values from the view of infants. This means that the object was re-defined from an infant's perspective because the number of possible accidents related to normal detached objects compared to those of special objects for babies remained the same. Moreover, the number of normal, daily-used objects were observed to be greater than that of baby products. This is necessary to provide different affordances to infants. The application shows concrete and specific examples with AR. Thus, it is possible to help to prevent "preventable" accidents.

4.2. AR Mobile Application

We measured the recognition rate and the accuracy of the vision-based system for AR applications to determine whether the AR systems functioned appropriately. The results showed differences between two AR systems depending on the conditions. We found no differences for distances in the 20–40 cm range; however, the tracking time was affected by the surface orientation. It took slightly longer to recognize the AR marker on the vertical surface than on the horizontal surface. Although this kind of problem is also found in Apple's ARKit, the subsequent update of ARKit will make it possible to recognize not only horizontal surfaces but also vertical surfaces as well. This platform is expected to become available in the summer of 2018 [45]. Therefore, we must re-examine and determine which platform is ideal to develop the application in the long run. Moreover, in real-world scenarios, marker tracking is sometimes unstable under indirect lighting. This suggests that the proposed application needs further improvements for everyday use. In summary, the marker-based AR system would be the most practical and adaptable application for the unique situation of a given user, considering the function and size of application at present. However, the number of objects in our everyday environments is increasing and its arrangements are also changing with time. A marker-less vision-based tracking system will be developed in the future using our on-going longitudinal recorded video resources.

We referred to Apple's human interface guidelines [46] to develop the application's interface. We focused on helping people understand and interact with the content when designing the interface. Considering that the evaluation of the AR system is important even though its conception is fairly recent, compared to developing AR studies. In 2008, Gabbard and Swan suggested the necessity to learn from user studies for user interface design, usability, and discovery, early in the development of emerging technologies such as AR [47,48].

4.3. Visualizing Home Environment with AR

We examined the validity of the aims and methods of this AR application through informal users interviews and user tests. Since 1995, when the first user-based experiments in AR literature were presented, usability studies have been published to address issues in three related areas, namely, perception, performance, and collaboration [2]. In this study, we focused on perception and performance.

The results revealed how users interact with virtual movies in an AR application in the real-world. Experts from 18 fields of research showed interests in the prototype and suggested improvements. This means that the visualization of affordances using AR technology is acceptable and the applications abilities are not limited to only parents. Furthermore, it indicated that our approach is actually suitable for interdisciplinary research with aging, architecture, computer science, design, education, healthcare, psychology, rehabilitation, and robotics. This may lead to the development of safety measures from positive and negative aspects in the everyday environment in the form of attractive and safe products, measurements and standard, AI-ready data, and smart homes.

Regarding data, almost all interviewees agreed that collecting longitudinal data in natural settings is worth exploring and considered as a good resource for finding solutions. This is one of our application's features. It is valued by professionals such as researchers, developers, and designers. However, this might also be a weakness because every AR animation requires significant time and effort to draw, because even a single movie requires at least 100 hand-drawings at the present. There is room for improvement and a more efficient method.

We fill our homes with objects; therefore, the application is aimed at enabling a re-examination of the way everyday objects are used, designed, and improved flexibly from the perspective of infants. Visualizing affordances related to an infant's home environment with an AR application would provide another structure to explore the creative processes of individuals who are responsible for coordinating the placement of objects at home, and the thinking behind possible solutions to the challenges we are facing currently to reduce preventable accidents and promote development for infants.

5. Conclusions and Future Directions

The work reported in this study represents an essential first step in adapting the concept of affordance theory to visualizing infants' perceptions and actions within the home environment and an attempt to illustrate the reality in a different perspective to caregivers of infants to prevent the accidents at home. Numerous incidents in the home environment result in injuries and even death of young children. Almost all the ordinary incidents of everyday life can pose danger to infants depending on the conditions. In this study, we demonstrate that AR technology is a promising and useful tool for visualizing invisible affordances of everyday objects. Our AR movies are based on the developmental level, or rather, the incidents anticipated owing to the developmental level of young children. To the scenery, we add the facts about infants and objects and enable caregivers to simulate and visualize other possible incidents. Thus, our app is different from injury reporting or warning apps.

Several limitations exist in this study. First, since we needed to add information on the developmental process of children, we conducted longitudinal observation; we could not determine how one specific fall becomes dangerous, because infants experienced thousands of steps and dozens of falls per day during natural locomotion [40]. It may be difficult to finish collecting data for a whole developmental period as that would delay analysis; however, we continue to collect data and increase the number of participants from year to year. Second, technical problems exist; adding AR markers and creating line drawing illustrations from videos are time-consuming. Therefore, an automatic process for generating line drawings from videos needs to be designed.

More broadly, from the results of interviews, we found that the prototype AR application has been easy and efficient for everyday use because smart devices and AR are common now. Our goal is to build an integrated system that will allow researchers, engineers, designers, and users worldwide to collaborate and contribute to the use of the proposed technology to promote the development of young children and prevent accidents in the home environment. We have started collecting data in one more country, as more evidence must be obtained by processing large amounts of relevant data in future works to generate further meaningful insights.

Conflict of Interest

The authors declare that there are no conflicts of interest associated with this manuscript.

Acknowledgment

We thank all the children and parents who participated in this research. We also thank Sachio Uchida, Tesuaki Baba and Shohan Hasan for their help, and João Barreiros and Rita Cordovil for their support in conducting the observations. This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 16K21263.

References

[1] M, Nishizaki, "Visualizing positive and negative affordances in infancy using mobile augmented reality" in Intelligent Systems Conference (IntelliSys), 1136-1140, 2017. DOI: 10.1109/IntelliSys.2017.8324272

[2] I. E. Sutherland, "A head-mounted three-dimensional display," in Proceedings of FJCC 1968, Thompson Books, Washington DC, 757–764, 1968.

[3] M. Billinghurst, A. Clark, G. Lee, "A Survey of Augmented Reality," Foundations and Trends in Human–Computer Interaction, 8(2-3), 73-272, 2015. http://dx.doi.org/10.1561/1100000049

[4] T. P. Caudell, D. W. Mizell, "Augmented reality: An application of heads-up display technology to manual manufacturing processes," In System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference, 2, 659–669. IEEE, 1992.

[5] S. Aukstakalnis, Practical augmented reality: a guide to the technologies, applications, and human factors for AR and VR, Addison-Wesley 2016.

[6] R. Azuma, M. Billinghurst, and G. Klinker, "Editorial: Special section on mobile augmented reality" Comput., Graph., 35(4), 7–8, 2011. https://doi.org/10.1016/j.cag.2011.04.010

[7] R. T. Azuma, "A survey of augmented reality" Presence-Teleop. Virt., **6**(4), 1892, 355–385, 1997.

[8] M. Mohring, C. Lessig, O. Bimber, "Video see-through AR on consumer cell-phones" in Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality, 252–253. IEEE Computer Society, 2004. DOI: 10.1109/ISMAR.2004.63

[9] O. Smørdal, G. Liestøl, and O. Erstad, "Exploring situated knowledge building using mobile augmented reality" in QWERTY 11, 1, 26–43, 2016.

[10] S. Rattanarungrot, M. White and B. Jackson, "The application of service orientation on a mobile AR platform — a museum scenario" 2015 Digital Heritage, Granada, 329-332, 2015. DOI: 10.1109/DigitalHeritage.2015.7413894

[11] D. Wagner, and D. Schmalstieg, "Making augmented reality practical on mobile phones, part 1" IEEE. Comput. Graph., 29(3), 12–15, 2009. http://doi.ieeecomputersociety.org/10.1109/MCG.2009.46

[12] D. Beier, R. Billert, B. Bruderlin, D. Stichling and B. Kleinjohann, "Marker-less vision based tracking for mobile augmented reality" The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings, 258-259, 2003. DOI: 10.1109/ISMAR.2003.1240709

[13] SUPERDATA Games & Interactive Media Intelligence, https://www.superdataresearch.com/us-digital-games-market/, last accessed 2018/08/13.

[14] Digi-Capital Homepage, https://www.digi-capital.com/news/20170/01/after-mixed-year-mobile-ar-to-drive-108-billion-vrar-market-by-2021/, last accessed 2018/05/03.

[15] The Ministry of Health, Labour and Welfare of Japan. List of Statistical Surveys. Vital Statistics 2016, https://www.e-stat.go.jp/en/stat-search/files?page=1&layout=datalist&tstat=000001028897&year=20160&month=0&tclass1=000001053058&tclass2=000001053061&tclass3=000001053066&result_back=1, last accessed 2018/05/03.

[16] The National Safety Council of US Homepage, http://www.rospa.com/homesafety/adviceandinformation/childsafety/accidents-to-children.aspx, last accessed 2018/05/03.

[17] Consumer Safety Unit. 24th Annual Report, Home Accident Surveillance System. London: Department of Trade and Industry, 2002.

[18] The U.S. Consumer Product Safety Commission, https://www.cpsc.gov/s3fs-public/5013.pdf, last accessed 2018/05/03.

[19] A. Carlsson, A. K. Dykes, A. Jansson, A. C. Bramhagen, "Mothers' awareness towards child injuries and injury prevention at home: an intervention study" BMC Research Notes, 9, 223, 2016. https://doi.org/10.1186/s13104-016-2031-5

[20] D. Kendrick, C. A. Mulvaney, L. Ye, T. Stevens, J. A. Mytton, S. Stewart-Brown, "Parenting interventions and the prevention of unintentional injuries in childhood: systematic review and meta-analysis" Cochrane Database System Review. 2013 Mar 28; (3):CD006020. Epub. https://doi.org/10.1111/j.1365-2214.2008.00849.x

[21] B. A. Morrongiello, S. Kiriakou, "Mothers' home-safety practices for preventing six types of childhood injuries: what do they do, and why?" J. Pediatr. Psychol., 29(4), 285–97, 2004.

[22] A. Carlsson, A. K. Dykes, "Precautions taken by mothers to prevent burn and scald injuries to young children at home: an intervention study" Scand. J. Public Health., 39(5), 471-478, 2011. https://doi.org/10.1177/1403494811405094

[23] SAFE KIDS WORLDWIDE, https://www.safekids.org/safetytips/field_venues/home, last accessed 2018/08/13.

[24] Japan Pediatric Society, Injury Alert and Follow-up report http://www.jpeds.or.jp/modules/injuryalert/, last accessed 2018/08/13.

[25] K. Kitamura, Y. Nishida, Y. Motomura, H. Mizoguchi, "Children Unintentional Injury Visualization System Based on Behavior Model and Injury Data" The 2008 International Conference on Modeling, Simulation and Visualization Methods (MSV'08), 2008.

[26] K. Kitamura, Y. Nishida, Y. Motomura, T. Yamanaka, H. Mizoguchi, "Web Content Service for Childhood Injury Prevention and Safety Promotion," in Proceedings of the 9th World Conference on Injury prevention and Safety Promotion, 270, 2008.

[27] Y. Nishida, Y. Motomura, K. Kitamura, T. Yamanaka, "Representation and Statistical Analysis of Childhood Injury by Bodygraphic Information System" Proc. of the 10th International Conference on GeoComputation. 2009.

[28] J. J. Gibson, "The Senses Considered as Perceptual Systems, Houghton. Mifflin Company, Boston, 1966.

[29] J. J. Gibson, "The ecological approach to visual perception," Boston: Houghton Mifflin, 1979.

[30] K. E. Adolph, "A psychophysical assessment of toddlers' abi1ity to cope with slopes" J. Exp. Psychol. Human, 21, 734–750, 1995.

[31] E. J. Gibson and R. D. Walk, "The "visual cliff" Scientific American, 202, 67–71, 1960.

[32] T. Stoffregen, "Affordances and events" Ecol. Psycol., 12, 1–28, 2010. https://doi.org/10.1207/S15326969ECO1201_1

[33] M. Turvey, "Affordances and prospective control: An outline of the ontology" Ecol. Psycol., 4, 173–187, 1992. https://doi.org/10.1207/s15326969eco0403_3

[34] W. H. Warren, "Perceiving affordances: Visual guidance of stair climbing" J. Exp. Psychol. Human, 10, 683–703, 1984.

[35] W. Gaver, "What in the world do we hear? An ecological approach to auditory source perception" Ecol. Psycol., 5, 1–31, 1993a. https://doi.org/10.1207/s15326969eco0501_1

[36] W. Gaver, "How do we hear in the world? Explorations in ecological acoustics" Ecol. Psycol., 5, 285–313, 1993b. https://doi.org/10.1207/s15326969eco0504_2

[37] L. S. Mark, "Perceiving the preferred critical boundary for an affordance," in Studies in perception and action III, B. G. Bardy, R. J. Bootsma, and Y. Guiard (Eds.), Mahwah, NJ: Lawrence Erlbaum Associates, 1995.

[38] L. S. Mark, K. Nemeth, D. Gardner, M. J. Dainoff, J. Paasche, M. Duffy, and K. Grandt, "Postural dynamics and the preferred critical boundary for visually guided reaching" J. Exp. Psychol. Human, **23**(5), 1365–1379, 1997.

[39] W. H. Warren, and S. Whang, "Visual guidance of walking through apertures: Body-scaled information for affordances" J. Exp. Psychol. Human, **13**(3), 371–384, 1987.

[40] K. E. Adolph, W. G. Cole, M. Komati, J. S. Garciaguirre, D.Badaly, J. M.Lingeman, G. L. Y.Chan, R. B. Sotsky, "How do you learn to walk? Thousands of steps and dozens of falls per day" Psychol. Sci., 23, 1387-1394, 2012. https://doi.org/10.1177/0956797612446346

[41] W. G. Cole, S. R. Robinson, & K. E. Adolph, "Bouts of steps: The organization of infant exploration" Dev. Psychobiol., 58, 341-354, 2016. https://doi.org/10.1002/dev.21374

[42] J. M. Franchak, and K. E. Adolph, "Visually guided navigation: Head-mounted eye-tracking of natural locomotion in children and adults" Vision Res., 50, 2766-2774, 2010. https://doi.org/10.1016/j.visres.2010.09.024

[43] Datavyu, http://www.datavyu.org/ last accessed 2018/10/10.

[44] Consumer Affairs Agency, Government of Japan, "Policy of children's accidents prevention in 2017" http://www.caa.go.jp/policies/policy/consumer_safety/child/children_accident_prevention/pdf/children_accident_prevention_180328_0004.pdf, last accessed 2018/08/13.

[45] Apple, Newsroom https://www.apple.com/jp/newsroom/2018/06/apple-unveils-arkit-2/ last accessed 2018/08/13.

[46] Apple, human-interface-guidelines, https://developer.apple.com/design/human-interface-guidelines/ios/overview/themes/ , last accessed 2018/08/05.

[47] J. E. Swan J. L. Gabbard. "Survey of user-based experimentation in augmented reality" in Proceedings of 1st International Conference on Virtual Reality, 1–9, 2005.

[48] J. L. Gabbard and J. E. Swan, "Usability engineering for augmented reality: Employing user-based studies to inform design" Vis. Comput. Graph., IEEE, 14(3), 513–525, 2008. http://doi.ieeecomputersociety.org/10.1109/TVCG.2008.24

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 3, 80-91 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Towards Process Standardization for Requirements Analysis of Agent-Based Systems

Khaled Slhoub*, Marco Carvalho

Florida Institute of Technology, School of Computing, Florida, 32901, United States

ARTICLEINFO

Article history:
Received: 18 April, 2018
Accepted: 11 May, 2018
Online: 31 May, 2018

Keywords:
IEEE Std 830
Agent-Oriented Specification
Agent-Based Systems
Multi-Agent Systems
MAS
ABS
Software Engineering
Software Requirements Specification
SRS
Software Standards
Software Quality

ABSTRACT

The development of agent-based systems is negatively impacted by the lack of process standardization across the major development phases, such as the requirements analysis phase. This issue creates a key barrier for agent technology stockholders regarding comprehending and analyzing complexity associated with these systems specifications. Instead, such fundamental low-level infrastructure is loosely attended to in an ad-hoc fashion, and important aspects of requirements analysis are often neglected altogether. The IEEE Std 830 model is a recommended practice aimed at describing how to write better quality requirement specifications of conventional software. Knowing that agent-based computing is a natural and logical evolution of the conventional approaches to software development, we believe that the requirements phase in agent-based systems can benefit from applying the IEEE Std 830 model which results in high-quality and more accepted artifacts. This article provides a criteria-based evaluation that is derived from the software engineering body of knowledge guide to assessing the adoption degree of agent-oriented methodologies to software requirements standards. Then, it proposes a model-driven approach to restructuring and extending the IEEE Std 830-2009 standard model to specify requirements of agent-based systems. To evaluate the applicability and usefulness of the updated model, we design a research study that allows practicing the model with simple real-world problem scenarios and conducting a summative survey to solicit feedback on the model usages.

1 Introduction

This article is an extension of work initially presented in 2017 at the IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) [1].

Agent-based systems (ABS) have proliferated in recent years into what is now one of the most active research areas in computing. This intensity of interest is increased not only because these complex systems impose a new promising means of developing software, but also because the agent community has become aware of the necessity to cast ABS as a software engineering paradigm (SE). This becomes important in order to assist ABS development to be more formal and efficient. Agent-Oriented Software Engineering (AOSE) is a new independent SE mainstream that aims to either extend or adapt existing SE methodologies to facilitate and improve the ABS complex development. SE practices have become a vital prerequisite of running successful software products and have been extensively used for decades to support the conventional ways of building software, such as object-oriented development which is currently the most popular programming paradigm [2].

Like other conventional software, software requirements specification (SRS) is an essential aspect in agent-based systems, and numerous agent-oriented methodologies demand it as an initial primary phase in their development process life-cycles. SRS describes how an agent system is expected to behave and extends the re-

*Corresponding Author: Khaled Slhoub, Email: kslhoub2014@my.fit.edu

https://dx.doi.org/10.25046/aj030311

quirements analysis phase to in-depth detail, detailed blueprints, used extensively by different stakeholders of agent systems. However, the complexity associated with ABS development and its target application domains often drives the complexity of ABS requirements phase. This results in poor quality SRS artifacts that have a genuine and significant negative impact on the entire ABS development life-cycle and create further technical problems to the stakeholders. With the emergence of industry willingness for agent systems, there is a real demand to develop a high-quality agent-oriented specification in such a way that makes ABS development easier, more formal, more disciplined, and more accepted industry-wide.

Over the past three decades, software standards have played a key role in improving quality of conventional software development. For instance, SE practices like the UML modeling language and IEEE/ISO standards have been widely used in the software industry to formalize the entire software development life-cycle. More specifically, the International Recommended Practice for Software Requirements Specifications model (IEEE Std 830-2009) has been commonly employed to help developers toward constructing high-quality and well-organized SRS artifacts [3]. Agent technology as another approach for software development can undoubtedly benefit from the IEEE Std 830-2009 model to standardize the SRS process which will mitigate the complexity of the requirements analysis and, as a result, produce more accepted SRS artifact. To do so, the IEEE Std 830-2009 model needs to be reconstructed and updated with new additional extensions to make it more suitable to handle the natural complexity of ABS and their target application domains.

In the original work, the intent was to restructure the IEEE Std 830-2009 model in such a way that makes it more suitable to handle ABS requirement specifications. We first explored the requirement phases in different well-known AOSE methodologies, described in [4], [5], and hence we identified five data models which need to be specified and incorporated into the ABS analysis process. Then, we checked the suitability of these data models according to the original IEEE Std 830-2009 sections and proposed rewriting some subsections, extending others, and adding more new extensions to the IEEE Std 830-2009 model to complement the process standardization in the ABS requirement phase. In this extended article, we explain in detail evaluation criteria used to determine the coverage degree of the agent-oriented methodologies with respect to the requirement standards. Also, we conduct an external evaluation study by asking participants to go through the updated model sections using a few simple real-world problem scenarios, and then take an online survey to provide feedback on their experience using the updated model.

The rest of this paper is organized as follows: section 2 gives the necessary overview of software agent concepts and goals, the agent-oriented specification process with some related work, and the structure of the original IEEE Std 830-2009 standard model along with some related work. Section 3 presents an evaluation matrix used to asses the coverage degree of agent-oriented methodologies with respect to requirement standards. Section 4 describes our proposed approach to restructuring the IEEE Std 830-2009 standard model in such a way that makes it more capable to address agent-oriented specification. Section 5 summarizes our evaluation method and results for the updated IEEE Std 830-200 model. 6 concludes.

2 Background and Related Work

2.1 Agent-Based Systems: What and Why

Over the past two decades, computing systems have become more complex. There are several reasons behind such complexity. First, numerous computing systems are now becoming concurrent and distributed. They are often configured and distributed over multiple locations with the use of multiple networks, tens of servers, and hundreds of different machines. Such systems also comprise a considerable number of software applications that extensively communicate with one another. Second, many computing systems operate in dynamic environments which are flexible enough to permit unpredictable changes to their contents and settings. Third, many computing systems now need to be continuously active and to provide services on an on-going basis [6]. Indeed, the explosive growth of the web and mobile technologies present a twofold problem in terms of the vast availability, wide variety, and heterogeneity of datasets [7]. Moreover, there has been a natural tendency to rely more on technology to resolve certain classes of complex real-world problems. People want computer systems to do more for them and, if possible, substitute them for performing costly and complex tasks. More than this, consumers demand systems to operate without their intervention and take the initiative when necessary. To do so, people build software components, known as agents, to act on behalf of users to achieve specific delegated goals. The intent is to develop sophisticated agents that are "capable of human-like reasoning and acting" [8], [9].

As a result of the above demands, new software characteristics and expectations have arisen and made current software development paradigms struggling to handle such complex development. These demands have also triggered a need to focus more on developing complex, real-time and distributed intelligent systems. Thus, this has prompted researchers to look for some effective alternative approaches to address the new demands of software development. Currently, agent-based computing is becoming the most promising development approach for handling such issues [5], [10], [11], [12], [13].

A software agent, another popular term used in-

terchangeably with an agent in the agent community, could be defined in several different ways depending on its usage and dimensions, we came across at least 12 formal definitions for the software agent. To more clearly understand the software agent, we can think of it as a role-playing smart or active software object that can interact with other agents object despite the fact that "agents are not simply objects by another name" [14]. Wooldridge highlights that "agents are simply software components that must be designed and implemented in much the same way that other software components are. However, AI techniques are often the most appropriate way of building agents" [14]. Jennings also supports this by stating that "an agent is an encapsulated computer system that is situated in some environment, and that is capable of flexible, autonomous action in that environment in order to meet its designed objectives" [15]. In fact, researchers consider agent-oriented development as a natural and logical evolution of the current software development paradigms like structured and object-orientated approaches as it presents a higher-level of abstraction and encapsulation [11], [16].

Actually, agent orientation represents not only a higher-level of abstraction and encapsulation that is more flexible than the other prior programming paradigms but also offers some unique dimensions or characteristics of agenthood such as **autonomy**, capability of deciding for itself how best to achieve delegated goals; **intelligence**, capacity of learning, reasoning and making educated decisions; and **socialability**, capability of interacting and coordinating with other agents. Such characteristics allow software agents to simulate human-like reasoning capability and to act rationally and flexibly to attain delegated goals. Yu highlights that "Agent orientation is emerging as a powerful new paradigm for computing" [17]. Jennings also states that "agent-oriented approaches can significantly enhance our ability to model, design and build complex, distributed software systems" [15].

Despite the success of several single-agent systems, a software agent as a computational entity rarely functions in isolation [18]. Any individual agent is primarily restricted by the limitation of its knowledge, resources, and potential which make it often useless on its own [15], [19]. With the current complex real-world problems, it is necessary to develop capable, intelligent distributed systems that employ a group of individual agents to resolve a combination of issues. Multi-agent systems (MAS) are constructed around this concept of embedding groups of agents that collaborate with one another, and, to do so, these individual agents require the ability to cooperate, coordinate, and negotiate with each other, much as people do. One broad definition for MAS is "a collection of interacting agents which are highly autonomous, situated and interactive software components. They autonomously sense their environment and respond accordingly. Coordination and cooperation between agents that possess diverse knowledge and capabilities facilitate the achievement of global goals that cannot be otherwise achieved by a single agent working in isolation" [20]. As shown in Figure 1, software agents in MAS are organized into teams or groups, each side with specific tasks to perform. The agent in its squad strategically partners with other agents and efficiently cooperates, coordinates, negotiates, and competes to fulfill the best outcome for itself and its team. Such collective interactions among the groups result in pursuing the entire delegated system goals which are beyond the limits of a single agent to accomplish [15]. In a study carried out in 2014 [21], Muller and Fischer investigated and analyzed 152 agent applications all over the world created by parties from 21 countries. They were able to demonstrate that MAS have attracted much more attention than other agent type applications (125 MAS applications corresponding to 82 % of all agent applications in the study).

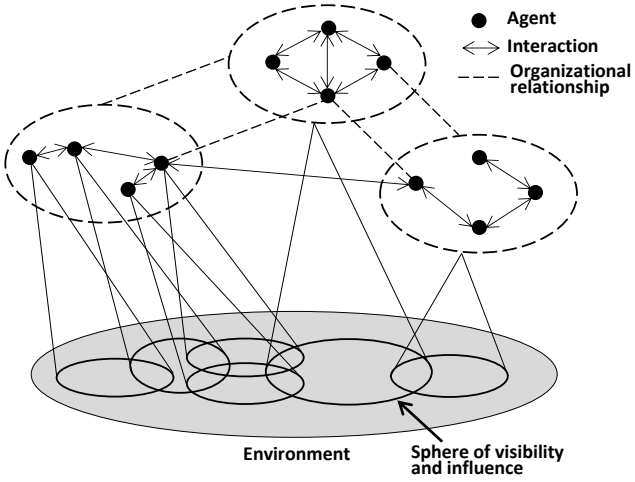


Figure 1: High-level structure of agent-based systems

2.2 The Requirements Analysis Process in Agent-Based Systems

The requirements analysis process is a significant development phase in large and complex conventional software. It is a description that leads to an understanding of the system and its structure prior to implementation, provides the foundation for the rest of development phases like design and testing, and acts as a reference point for the system stakeholders. ABS, which are structured based on communicating software agents, are not an exception as various existing agent-oriented development methodologies identify the requirements analysis phase as the initial phase that developers of agent systems need to fulfill. However, unlike conventional software, the requirements phase in agent systems requires to cover more additional complex details of software agents, their roles, their dynamics interactions, their cooperation patterns, and internal behaviors combined with other information of their surrounded physical/virtual and operating environments, and the target application domain.

To deal with such complexity, Ferber et al. de-

scribe the ABS requirements from the organizational point of view by using the notions of agent, group, and role models. Based on this structure, requirements are classified into four types: single role behavior requirements, intragroup interaction requirements, intragroup communication successfulness requirements, and intergroup interaction requirements [22]. Similarly, Hilaire et al. aim to analyze complex agent communications by proposing an organizational model centered around agent interaction patterns that are composed of roles [23]. Instead, Herlea et al. introduce three levels of abstraction for the ABS requirement: requirements for the multi-agent system as a whole, requirements for an agent within the system, and requirements for a component within an agent [24]. Miller et al. indicate the need for adequate elicitation methods of agent-oriented requirements to mitigate complexity. They propose an elicitation approach that provides a list of questions to be answered by stakeholders and then map the answers to the ABS requirements [25].

Most of the existing agent-oriented development methodologies are derived originally from some software engineering processes models used in supporting the development of conventional software such as waterfall, evolutionary, and RUP [26], [27]. Such process models give significant attention to the requirements engineering phase to delineate the complexity issue and avoid situations like poor time and budget estimations, unpredictable outcomes, and customer dissatisfaction [28]. To do so, the phase is often supported by sets of standards, diagrammatic notations, and various CASE tools. Similarly, agent-based systems are basically software components that need to be developed in much the same ways of developing conventional software [15], [16]. Thus, these systems can also benefit from the existing SE practice in improving the requirements analysis phase. One essential improvement, for instance, is to apply different types of software documentation standards to increase the quality of software development. IEEE/ISO/IEC standards have been widely used to support conventional software development. For example, the IEEE Std 830 documentation model is widely used to support software requirements specification and can also play a key role not only in improving and formalizing the requirement analysis process in agent systems but also in giving a better signal to the industry to better understand the ABS concepts and realize their benefits.

2.3 The IEEE Std 830-2009 Standard Practice Model

Very little work has been reported in the literature to standardize and formalize agent-based systems development. For example, Foundation for Intelligent Physical Agents (FIPA) [29], accepted as a standard committee of the IEEE Computer Society in 2005, is proposed to facilitate agent aspects like agent platform, protocols, and communication languages. However, FIPA standard templates are large, complex, and hard to follow. Such templates are considered more appropriate to address the agent systems design phase but not the requirements phase [29], [30]. Agent modeling languages are another example proposed by extending the UML metamodel to meet the structural aspects of agent systems like TAO [31], AML [32] and Agent UML [33]. Such modeling languages are, however, restricted by object-oriented concepts and cannot tackle the required additional complex details of software agents mentioned in section 2.2.

Meanwhile, the International Recommended Practice for Software Requirements Specifications model (IEEE Std 830-2009) demonstrates a simple generic model to guide developers in the requirements analysis process [3]. IEEE Std 830 was specifically developed to standardize the process of software requirements specifications (SRS) in conventional software, also known as the high-level design of the system, and to drive developers towards producing high-quality and well-organized SRS artifacts. In addition, it aims to establish the basis for agreement between customers and suppliers of as to what the final software product should deliver; reduce the development time and effort; estimate efficiently project cost and scheduling; facilitate the smooth transactions between projects stages; ensure continuity of work; enhance software maintenance and reuse; and provide the baseline for testing [3]. The IEEE Std 830 model frequently serves as a reference for developers during software development processes and as a contract between project customers and suppliers. In fact, it has been widely used in software industry especially in supporting the object-oriented development, and it is recognized according to IEEE as the most popular and universal standard among all other IEEE standards with 50,925 full-text views in 20 years [3], [34]. As shown in Figure 2, the IEEE Std 830 model is divided into four main sections composed of subsections that support achieving specific objectives. Refer to [3] for more information.

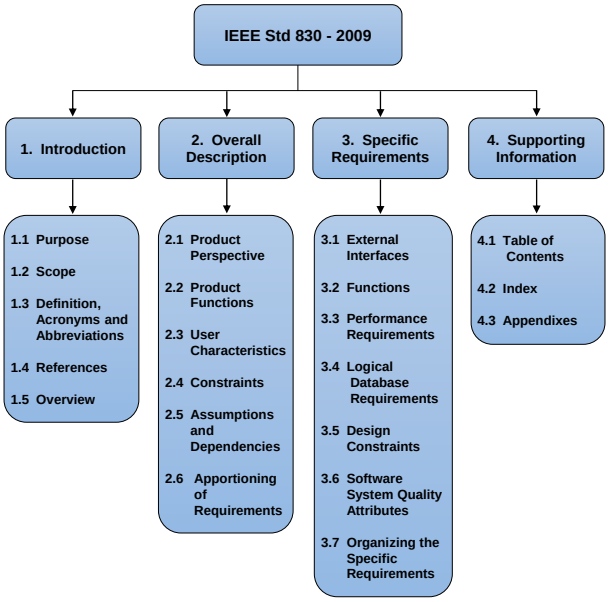


Figure 2: The IEEE Std 830-2009 model structure

3 Evaluating the Adoption of ABS Requirements to Software Standards

ABS development methodologies pay a great deal of attention to the requirements analysis process and identify it as an initial phase that ABS developers need to fulfill. We deeply explored the requirement phases in a large number of agent-oriented methodologies searching for evidence of implementing sound software standards. We wanted to determine how well the requirement phases adopt and practice software standards. To do so, we provided a simple criteria-based evaluation matrix that is basically derived from the software engineering body of knowledge (SWEBOK-V3.0) guide. SWEBOK-V3.0 is a well-defined, structured set of SE attributes used to deliver one of the essential benchmarks underlying the software engineering baseline [35]. Then, we used the evaluation matrix as a checklist to assess the coverage degree of an agent-oriented methodology with respect to requirement standards. Table 1 illustrates the evaluation matrix checklist combined with a set of agent-oriented development methodologies that were inspected for requirements standardization.

As shown in Table 1, the evaluation matrix was applied to more than 20 methodologies defined in [4], [5]. These methodologies were selected in light of the ongoing flow of publications, the remarkable impact on the AOSE community, the ease of understanding, and the profusion of supported guidance and tools. Despite the fact that all the inspected methodologies describe requirements analysis in sufficient detail, we found that the standards practice is often negated or rarely mentioned in these processes. In fact, to the best of our knowledge, no well-structured process standardization has been proposed and incorporated into any requirements analysis phase in the existing ABS development methodologies. We argue that the lack of process standardization in requirements analysis is linked to several AOSE challenges, such as immature ABS development methodologies, the absence of unified ABS methodologies and notation, and weak industrial acceptance of the AOSE paradigm.

4 Using the IEEE Std 830-2009 Model to Standardize Agent-Oriented Specification

Agent-based systems can be seen as a natural evolution from the object orientation paradigm in that they offer a higher-level of abstraction and encapsulation [11], [15], [16]. ABS are basically software components combined with additional dimensions or characteristics like autonomy, intelligence, and socialability that make ABS more flexible and suitable to address certain classes of complex real-world problems. ABS need to be developed in much the same ways as other conventional software components are developed. Thus, adopting the IEEE Std 830-2009 model, used with conventional software, in the requirement analysis phase of ABS development could become a core motif in promoting the entire ABS development quality and also map the way for the industry to accept and recognize the benefits of agent technology. To do so, The IEEE Std 830-2009 model needs to be restructured and extended to cover all agent-oriented specifications.

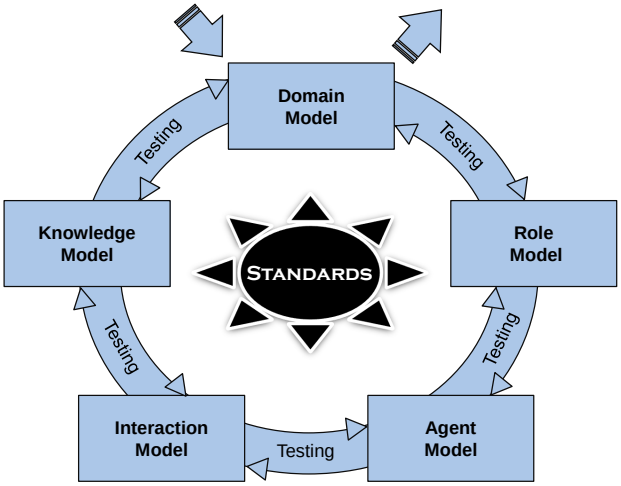


Figure 3: The five data sets required to meet agent-oriented specification [1]

As mentioned in section 2.3, the IEEE Std 830-2009 model was designed to provide recommended approaches for handling the software requirements specification. ABS can benefit from this standard model to

Evaluation Factor	Agent-Oriented Methodology
define clearly a process to elicitation, analysis, specification, and validation of requirements	PASSI, MaSE, Prometheus, Tropos, MAS-CommonKADS, Gaia, ADELFE, MESSAGE, INGENIAS, AOR, SODA, DESIRE, Agent Factory, MADE, AOSM, Agent OPEN, FIPA, ODAM, MASSIVE, Roadmap
monitor the quality and improvement of the requirements process like using quality standards and metrics	
establish software and system requirements documents	
use quality indicators to improve SRS	
provide an information basis for transferring work and software enhancement	
review the deviation from standard practice	

Table 1: The evaluation matrix used to determine the adoption degree of ABS requirements to standards

shape and formalize the requirements analysis process in such a way that results in more accepted specification document. Thus, we attempt in this section to specify ABS requirements by means of utilizing the IEEE Std 830-2009 model and proposing several new extensions to it. Accordingly, we first identified the main data models in the SRS process of ABS by investigating the analysis processes in several different AOSE development methodologies such as PASSI [36], ADELFE [37], MaSE [38], Gaia [39], and Tropos [40]. Then, we checked the fitness of these data models according to the IEEE Std 830-2009 model. Consequently, we were able to identify the data models that are not covered by the original IEEE Std 830-2009 model and proposed rewriting some subsections, extending others, and adding more new extensions to the model that can facilitate the agent-oriented specification. Figure 3 illustrates the identified 5 data models that carry the required information for specifying the ABS requirements. The 5 data models are described in the following list:

1. **The Domain model** is concerned with the analysis of a specific real world that an agent-based system is intended to work in and interact with. This model consists of three types of information:
    - *Operating Environment* includes descriptions of the general characteristics of the operating environment of ABS. Such environment provides the necessary infrastructure in which agents deployed and live. It also manages these agents while performing their tasks and delivers the most current information about them. Moreover, it is responsible for providing the structure of the agent communication language and the agent semantic language.
    - *Physical/Virtual Environment* includes descriptions of the world that software agents perceive and interact with. ABS are often executed in challenging environments that are dynamic, unpredictable, and unreliable. Such environments could be virtual like the web world or physical like a self-driving car and a train control system.
    - *Application Domain* includes information about the field in which the system operates. Such information may involve certain domain terminology or reference to domain concepts and policies.
2. **The Role model** is concerned with describing the ABS requirements in terms of agent tasks or services. Agents are the core actors that are responsible for performing tasks and achieving delegated goals. The descriptions may include information about the role responsibilities, prerequisites, and constraints.
3. **The Agent model** is concerned with identifying and classifying agent types that are instantiated later at ABS run-time, and it is directly linked to the Role model. The Agent model should also include descriptions of characteristics for every agent type which may include at least descriptions of the essential characteristics of agents such as autonomy, reactiveness, proactiveness, and socialability [15], [16].
4. **The Interaction model** is concerned with describing the agents behavior in terms of cooperating, collaborating, negotiating with users, system resources, and other agents. It should also specify the agents messaging protocols and interaction patterns along with the use of ontologies or conceptual means for describing the message contents. For instance, an agent can request another agent to perform some action, and the other agent may refuse or accept the request. It may also confirm to other agents that given proposition is true.
5. **The Knowledge model** is concerned with describing a repository of knowledge that agents may use to provide explanations, recovery information, or optimize their performance. For instance, some agent types, like reflex agents and learning agents, require some particular rules in order perform actions, so such rules need to be specified and stored in the system. The Knowledge model, also, should briefly describe agent architectures which underlying the concepts for building rational agents and their characteristics. Such information is important to design any agent-based system later. For instance, Belief-Desires-Intentions (BDI) architecture is probably the best-known agent architecture that agents rely on to reason about their actions [41].

After identifying the five main data models for agent-oriented specification, we carefully reviewed the four main sections of the original IEEE Std 830-2009 model to check the suitability of the 5 data models according to these sections. As a result, we were able to propose adding more new subsections, rewriting others, and extending other subsections to formulate the ABS requirements specification. Figure 4 illustrates the updated structure of the IEEE Std 830-2009 model by highlighting the new and edited subsections. Also, full descriptions are provided for all new information added to the updated model as follows:

1. The first section **"Introduction"** contains subsections that provide an overview of the entire SRS document, so we consider this information suitable for the five data models of ABS [3]. We only suggest the following change in the following subsection:
    - An existing subsection (1.5 Overview): this subsection should provide a brief description of the contents of the rest of SRS document, so it should also include a description

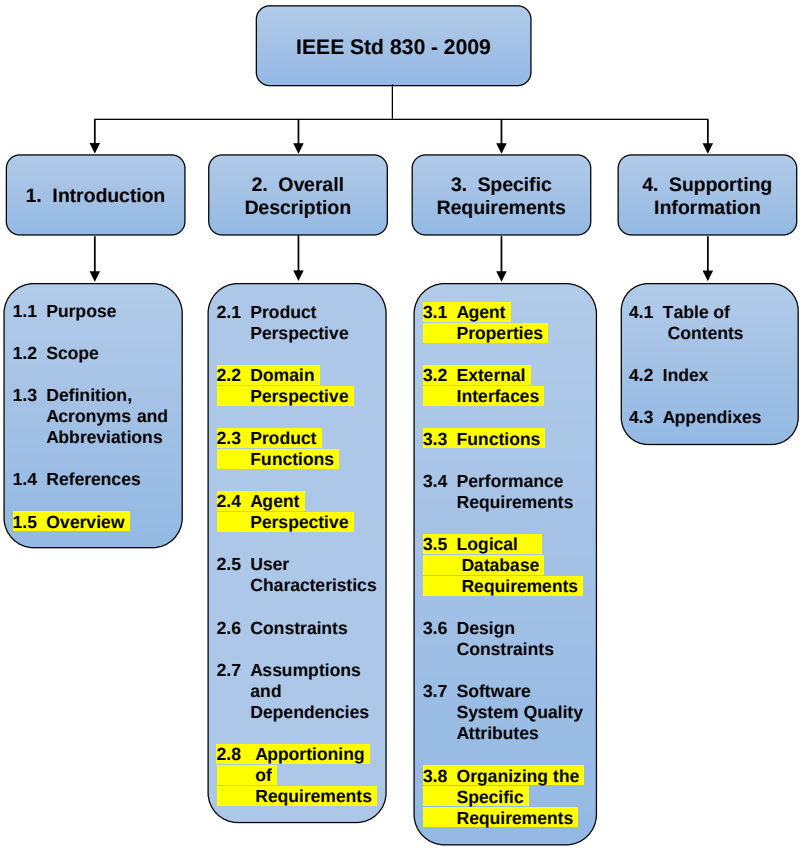


Figure 4: The updated structure of the IEEE Std 830-2009 model [1]

of how analyses of the five ABS data models are organized and presented in the SRS model.

2. The second section **"Overall Description"** contains subsections that describe the general factors that could affect the product and its requirements specifications [3]. We suggest the following changes to this section:

   - A new defined subsection (2.2 Domain Perspective): based on the provided description of this section, the *Domain model* should be described after (2.1 Product Perspective) for an organizational purpose. The domain application often has a great influence on ABS, ABS are considered as environment-driven business development. So, the potential features of the three sections of the *Domain model* are described in this subsection.
   - An existing subsection (2.2 Product Functions): this subsection is renumbered to subsection (2.3), and it is described in terms of both: the *Role model* and the user functions. Thus, it should provide a summary of the key roles that agents in ABS will perform. Notice that this only include a high-level summary of the *Role model* that defines the major functional roles for each listed agent.
   - A new defined subsection (2.4 Agent Perspective): based on the provided description of this section, the *Agent model* should be summarized in section 2. It should include a description that lists and defines the potential software agent categories and their main characteristics.
   - An existing subsection (2.6 Apportioning of Requirements): this subsection is renumbered to subsection (2.8), and it should identify functional agent roles and user functions that may be delayed until future versions of the agent-based system.

3. The third section **"Specific Requirement"** contains subsections that describe in sufficient technical details all ABS requirements. This includes all functional roles concerning all inputs and outputs to/from the agent-based system and all required nonfunctional requirements [3]. We suggest the following changes to this section as follows:

   - A new defined subsection (3.1 Agent Properties): this subsection should extend the subsection (2.4 Agent Perspective). It is concerned with agent properties/ dimensions to a level of sufficient detail that will help ABS designers to construct the internal states for every agent type in ABS.
   - An existing subsection (3.1 External Interfaces): this subsection is renumbered to subsection (3.2). One goal of the third section is

to provide detailed descriptions of software interfaces. So, the *Interaction model* should be described in this subsection in such a way that presents communication mechanisms and messages between all agent types.

- An existing subsection (3.2 Functions): this subsection is renumbered to subsection (3.3). It is concerned with all ABS requirements specifications to a level of detail. This should extend the subsection (2.2 Product Functions) and include sufficient detailed specifications for every functional agent role and user functions. Also, this subsection should extend the subsection (2.4 Agent Perspective) and include sufficient detailed specifications of software agents. It is critical also to explicitly link listed proposed agents to their functional roles. An agent-role list is an effective technique to use here.
- An existing subsection (3.4 Logical Database Requirements): this subsection is renumbered to subsection (3.5). It is concerned with any information that is to be stored in a database. Thus, the *knowledge model* should be described in such a way that shows how some types of agents, like the reflex and learning agents, can use it. Also, the agent architectures should be briefly outlined in this subsection.
- An existing subsection (3.7 Organizing the Specific Requirements): this subsection is renumbered to subsection (3.8). The agent-oriented specification is detailed and tends to be extensive. So, it is important to give special attention to organizing this information to ensure clarity and understandability. This section should include an additional organization which organizing by agent class. Associated with each agent class is a set of agent roles and characteristics.

4. The fourth section **"Supporting Information"** contains subsections that make the SRS document easier to use, for example, using the table of contents, index, and appendixes [3]. We suggest no change to this section and consider this information suitable for the five data models of ABS.

5 Evaluation Analysis

In the original published paper, we conducted two types of evaluation to assess the utility and effectiveness of the proposed restructured IEEE Std 830 model: 1) we specified requirements of a small multi-agent system by using both the new proposed model and the original model, and then we compared the outcomes of both models. 2) we used our proposed model to specify requirements of some previous agent systems that were already specified by using the original model, and we analyzed the outcomes of the different models based on the SRS quality attributes described in the IEEE Std 830-2009 model. Refer to [1] for more information about both assessments.

This article summarizes results from an independent and external research study that was conducted to assess the suitability and validity of the updated IEEE Std 830. The aim of the study was to determine whether the updated model satisfies ABS requirements specification and to solicit feedback on any possible missing details or weaknesses in the model structure. Several experts in software engineering and agent-oriented systems were invited to participate in the research study. They were asked to go through the proposed model sections with a few simple real-world problem scenarios (requirements specification of a simple small agent-based system: *Item-Trading Agent-Based System (ITABS)* that enables users to create and employ agents to act on behalf of them to trade items). Then, the participants conduct an online survey to provide feedback on their experience using the proposed model in specifying ITABS. The survey consisted of 12 statements as shown in Table 2. These statements were organized into 3 sets to determine participants' acceptance of the proposed model, their satisfaction with using the various added and updated agent information to the model, and the quality of the proposed model. The participants were asked to indicate their level of agreement or disagreement with each statement. We were able to collect and analyze 13 responses from the participants as shown in the following charts.

In Figure 5, our concern was focused on whether or not the participants think that the proposed standards model is capable of specifying ABS requirements. The participants were all in favor of the model as no one disagreed with the statements S1 and S2.

In Figure 6, the participants provided their feedback about the added and updated agent information to the original IEEE Std 830 model. We believe that the existence of such information makes the IEEE Std 830-2009 model more suitable to handle the agent-oriented specification. The positive responses actually confirmed our belief as all participants agreed with the statements S3 and S4, and only 4 persons disagreed with the statement S5. The proposed model is required to briefly outline agent architectures which support the concepts for building rational agents and their characteristics. This information is essential to start any ABS design later, but it seems that some participants disagree with referring to implementation details, like referring to agent architectures in the requirement phase.

In Figure 7, the participants gave their opinions about the quality of the proposed model after using it in specifying small agent-based system requirements. They rated the model based on the IEEE seven quality attributes of good software requirements, described in the original IEEE Std 830 model, by indicating how well they feel that the requirements meet the quality

S1	The updated IEEE 830 model helps the developers to identify an agent-based system to be developed along with its requirements.
S2	The updated IEEE 830 model describes what the agent-based system will do and how it will be expected to perform.
S3	Information regarding domain perspective which includes application domain, operating environment, and physical/virtual environment needs to be specified in the IEEE 830 model and will be used to build an agent-based system.
S4	Information regarding agent classification, characteristics, functions, and communications needs to be specified in the IEEE 830 model and will be used to build an agent-based system.
S5	Information regarding agent architectures and agent repository of knowledge, if there's any, needs to be briefly outlined in the IEEE 830 model and will be used to build an agent-based system.
S6	The functional requirements described in the model were **unambiguous**: there was only one interpretation for every stated requirement.
S7	The functional requirements described in the model were **complete**: all significant requirements concerning function, performance, interfaces, design constraints, and quality factors were included.
S8	The functional requirements described in the model were **consistent**: no subset of individual requirements characteristics, actions, or terms conflicted.
S9	The functional requirements described in the model were **verifiable** (Testable): requirements were stated in concrete terms that can be verified through testing or other objective means.
S10	The functional requirements described in the model were **modifiable**: the specification was structured and annotated so that changes may be made easily, completely, and consistently.
S11	The functional requirements described in the model were **traceable**: the origin of each requirement was clear and the format should aid in tracing forward and backward.
S12	The functional requirements described in the model were **usable**: the specifications should be useful during development and in later identifying maintenance requirements and constraints.

Table 2: The online survey statements given to the participants to scale

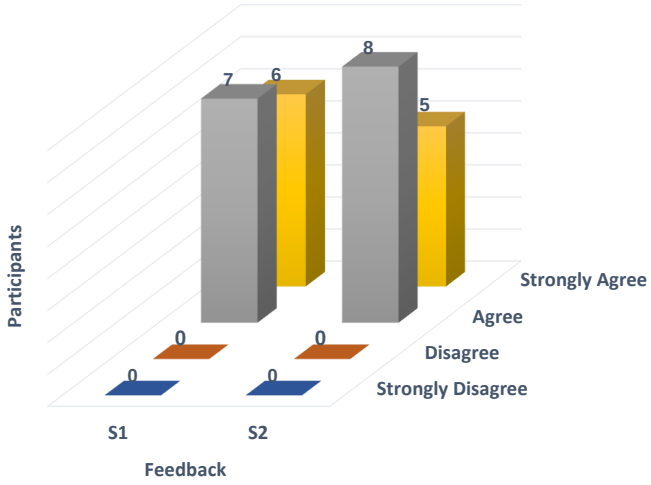


Figure 5: The degree of acceptance of the updated IEEE Std 830-2009 model

attributes on a scale from Strongly Agree to Strongly Disagree. Generally, the responses were positive with a range from 62% to 93% of the participants agreed or strongly agreed with each quality attribute applied to the updated model.

Figure 8 represents a straightforward view that only illustrates the summary of users' views who agreed with each quality attributes applied to the proposed standard model.

6 Conclusion

A strong demand desires to apply agent-based systems in complex and large real-world applications because of their capability to act rationally and flexibly to attain delegated goals. ABS development requires the inclusion of sound software engineering practices in order to be more formal, efficient, and adopted in the industry. Software standards as an SE key practice are widely applied in the software industry to support the entire software development life-cycle. Based on our evaluation applied to a large number of agent-oriented methodologies, we believe that no well-structured process standardization has been proposed and incorporated into any requirements analysis phase in the existing agent-oriented methodologies.This paper addresses the process standardization for requirements specification of agent-based systems by proposing a model-driven approach to restructure and update the IEEE Std 830-2009 model to make it more suitable to handle the ABS requirement specifications.

The updated IEEE Std 830-2009 model proposed adding new extensions and updating others in the orig-

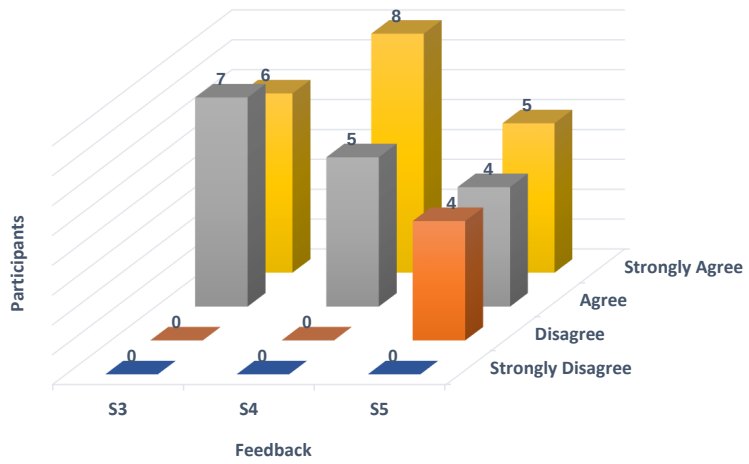


Figure 6: The degree of user satisfaction after applying the updated IEEE Std 830-2009 model

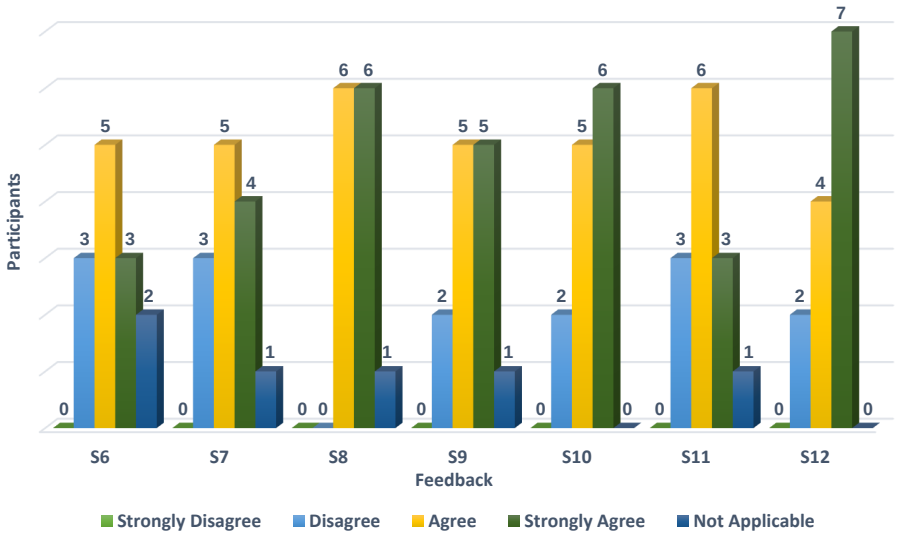


Figure 7: The summary of users' views regarding the given quality attributes as applied to the updated model

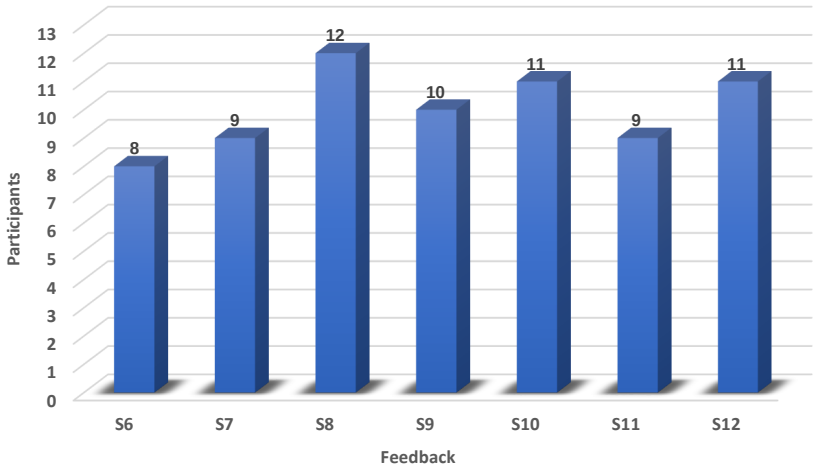


Figure 8: The summary of users' views who agreed with each quality attribute

inal model to complement the process standardization in ABS requirement specifications. We demonstrated the value of using the updated model by conducting several different types of evaluations. This paper describes an independent and external approach to evaluate the proposed model by asking expert participants to walk through the proposed model sections with a few simple real-world problem scenarios and then

take an online survey to provide feedback on their experience using the model. The feedback was very encouraging as all participants agreed that the proposed model was capable of handling the ABS requirements, and were satisfied with using the model along with all updated and added agent information to it. Also, The summary of users' views with respect to the given quality attributes as applied to the updated model were positive and reflected how well the agent-oriented specification in the updated model met these quality attributes.

In the future, we plan to extend the updated IEEE Std 830-2009 model to cover more specific sub-standards for every ABS data model we identified and for every new central extension we added. This can be done by studying what sub-standard information is required to be specified in every extension. For instance, what ABS sub-standard information in the *section 2.2 domain perspective* needs to be defined, or what sub-standard information should be included in the *section 3.1 Agent properties*.

Conflict of Interest The authors declare no conflict of interest.

References

[1] K. Slhoub, M. Carvalho, and W. Bond, "Recommended practices for the specification of multi-agent systems requirements," in *Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), 2017 IEEE 8th Annual*, pp. 179–185, IEEE, 2017. doi:10.1109/UEMCON.2017.8249021.

[2] B. Bauer and J. P. Müller, "Methodologies and modeling languages," *Agent-based Software Development*, pp. 77–131, 2004.

[3] IEEE, *IEEE Recommended Practice for Software Requirements Specifications*, vol. 2009. IEEE, 1998. doi:10.1109/IEEESTD.1998.88286.

[4] B. Henderson-Sellers, *Agent-oriented methodologies*. Idea Group Inc., 2005.

[5] O. Z. Akbari, "A survey of agent-oriented software engineering paradigm: Towards its industrial acceptance," *International Journal of Computer Engineering Research*, vol. 1, no. 2, pp. 14–28, 2010.

[6] F. Zambonelli, N. R. Jennings, and M. Wooldridge, "Developing multiagent systems: The Gaia methodology," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 12, no. 3, pp. 317–370, 2003. doi:10.1145/958961.958963.

[7] Q. H. Mahmoud, "Software Agents: Characteristics and Classification," *School of Computing Science, Simon Fraser University*, pp. 1–12, 2000.

[8] M. Wooldridge and N. R. Jennings, "Pitfalls of agent-oriented development," in *Proceedings of the second international conference on Autonomous agents*, pp. 385–391, ACM, 1998. doi:10.1145/280765.280867.

[9] A. Sloman, "What sort of architecture is required for a human-like agent?," in *Foundations of rational agency*, pp. 35–52, Springer, 1999. doi:10.1007/978-94-015-9204-8_3.

[10] F. Zambonelli and A. Omicini, "Challenges and research directions in agent-oriented software engineering," *Autonomous Agents and Multi-Agent Systems*, vol. 9, no. 3, pp. 253–283, 2004. doi:10.1023/B:AGNT.0000038028.66672.1e.

[11] L. Padgham and J. Thangarajah, "Agent Oriented Software Engineering: Why and How," *VNU Journal of Science: Natural Sciences and Technology*, vol. 27, no. 3, pp. 190–204, 2011.

[12] S. Maalal and M. Addou, "A new approach of designing Multi-Agent Systems With a practical sample," *CoRR*, vol. abs/1204.1, pp. 148–157, 2012.

[13] O. Shehory and A. Sturm, "A brief introduction to agents," in *Agent-Oriented Software Engineering: Reflections on Architectures, Methodologies, Languages, and Frameworks*, vol. 9783642544, pp. 3–11, Springer, 2014. doi:10.1007/978-3-642-54432-3_1.

[14] M. Wooldridge, "Agent-based software engineering," *IEEE Proceedings-software*, vol. 144, no. 1, pp. 26–37, 1997. doi:10.1049/ip-sen:19971026.

[15] N. R. Jennings, "On agent-based software engineering," *Artificial Intelligence*, vol. 117, no. 2, pp. 277–296, 2000. doi:10.1016/S0004-3702(99)00107-1.

[16] M. Wooldridge, *An Introduction to MultiAgent Systems*. Wiley, 2nd ed., 2009.

[17] E. Yu, "Agent-oriented modelling: Software versus the world," in *Artificial Intelligence and Bioinformatics*, vol. 2222, pp. 206–225, Springer, 2002. doi:10.1007/3-540-70657-7_14.

[18] Z. Ren and C. J. Anumba, "Multi-agent systems in construction-state of the art and prospects," *Automation in Construction*, vol. 13, no. 3, pp. 421–434, 2004. doi:10.1016/j.autcon.2003.12.002.

[19] P. Bogg, G. Beydoun, and G. Low, "When to use a multi-agent system?," vol. 5357 LNAI, pp. 98–108, Springer, 2008. doi:10.1007/978-3-540-89674-6_13.

[20] M. Dastani, "A survey of multi-agent programming languages and frameworks," in *Agent-Oriented Software Engineering*, vol. 9783642544, pp. 213–233, Springer, 2014. doi:10.1007/978-3-642-54432-3_11.

[21] J. P. Müller and K. Fischer, "Application impact of multi-agent systems and technologies: A survey," in *Agent-Oriented Software Engineering*, pp. 27–53, Springer, 2014. doi:10.1007/978-3-642-54432-3_3.

[22] J. Ferber, O. Gutknecht, C. M. Jonker, J. P. Müller, and J. Treur, "Organization models and behavioural requirements specification for multi-agent systems," in *Proceedings - 4th International Conference on MultiAgent Systems, ICMAS 2000*, pp. 387–388, IEEE, 2000. doi:10.1109/ICMAS.2000.858488.

[23] V. Hilaire, A. Koukam, P. Gruer, and J. J.-P. Müller, "Formal Specification and Prototyping of Multi-agent Systems," in *Engineering Societies in the Agents World SE - 9*, vol. 1972, pp. 114–127, Springer, 2000. doi:10.1007/3-540-44539-0_9.

[24] D. E. Herlea, C. M. Jonker, J. Treur, and N. J. E. Wijngaards, "Specification of bahavioural requirements within compositional multi-agent system design," vol. 1647, pp. 8–27, Springer, 1999. doi:10.1007/3-540-48437-X_2.

[25] T. Miller, B. Lu, L. Sterling, G. Beydoun, and K. Taveter, "Requirements elicitation and specification using the agent paradigm: The case study of an aircraft turnaround simulator," *IEEE Transactions on Software Engineering*, vol. 40, no. 10, pp. 1007–1024, 2014. doi:10.1109/TSE.2014.2339827.

[26] P. Giorgini and B. Henderson-Sellers, *Agent-Oriented Methodologies: an Introduction*. Idea Group Inc., 2005.

[27] L. Cernuzzi, M. Cossentino, and F. Zambonelli, "Process models for agent-based development," *Engineering Applications of Artificial Intelligence*, vol. 18, no. 2, pp. 205–222, 2005. doi:10.1016/j.engappai.2004.11.015.

[28] E. Ajith Jubilson, P. Joe Prathap, V. Vimal Khanna, P. Dhanavanthini, W. Vinil Dani, and A. Gunasekaran, "An empirical analysis of agent oriented methodologies by exploiting the life-cycle phases of each methodology," in *Advances in Intelligent Systems and Computing*, vol. 337, pp. 205–214, Springer, 2015.

[29] FIPA, "The standards organization for agents and multi-agent systems," 2005.

[30] P. Charlton, R. Cattoni, A. Potrich, and E. Mamdani, "Evaluating the FIPA standards and their role in achieving cooperation in multi-agent systems," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, vol. 00, pp. 1–10, IEEE, 2000. doi:10.1109/HICSS.2000.926996.

[31] V. T. Da Silva and C. J. P. De Lucena, "Modeling multi-agent systems," *Communications of the ACM*, vol. 50, no. 5, pp. 103–108, 2007. doi:10.1145/1230819.1241671.

[32] I. Trencansky and R. Cervenka, "Agent Modeling Language (AML): A comprehensive approach to modeling MAS," *INFORMATICA-LJUBLJANA-*, vol. 29, no. 4, pp. 391–400, 2005. doi:10.1.1.60.8902.

[33] B. Bauer, J. P. Müller, and J. Odell, "Agent UML: A formalism for specifying multiagent software systems," *International journal of software engineering and knowledge engineering*, vol. 11, no. 03, pp. 207–230, 2001. doi: 10.1142/S0218194001000517.

[34] A. Chikh and M. Aldayel, "Reengineering requirements specification based on IEEE 830 standard and traceability," in *Advances in Intelligent Systems and Computing*, vol. 275 AISC, pp. 211–227, Springer, 2014. doi:10.1007/978-3-319-05951-8_21.

[35] P. Bourque and R. E. Fairley, *Guide to the Software Engineering Body of Knowledge (SWEBOK(R)): Version 3.0*. Los Alamitos, CA, USA: IEEE Computer Society Press, 3rd ed., 2014.

[36] M. Cossentino and C. Potts, "PASSI: A process for specifying and implementing multi-agent systems using UML," *Retrieved October*, vol. 8, p. 2007, 2002.

[37] C. Bernon, M. Gleizes, S. Peyruqueou, and G. Picard, "ADELFE: A methodology for adaptive multi-agent systems engineering," in *International Workshop on Engineering Societies in the Agents World*, pp. 156–169, Springer Berlin Heidelberg, 2003. doi:10.1007/3-540-39173-8_12.

[38] M. Wood and S. Deloach, "An Overview of the Multiagent Systems Engineering Methodology," in *Proceedings - the First International Workshop on Agent-Oriented Software Engineering*, vol. 1957, pp. 207–221, Springer, 2001.

[39] M. Wooldridge, N. R. Jennings, and D. Kinny, "The Gaia methodology for agent-oriented analysis and design," *Autonomous Agents and multi-agent systems*, vol. 3, no. 3, pp. 285–312, 2000. doi:10.1023/A:1010071910869.

[40] P. Bresciani, A. Perini, P. Giorgini, F. Giunchiglia, and J. Mylopoulos, "Tropos: An agent-oriented software development methodology," *Autonomous Agents and Multi-Agent Systems*, vol. 8, no. 3, pp. 203–236, 2004. doi:10.1023/B:AGNT.0000018806.20944.ef.

[41] M. Georgeff, B. Pell, M. Pollack, M. Tambe, and M. Wooldridge, "The Belief-Desire-Intention Model of Agency," in *International Workshop on Agent Theories, Architectures, and Languages*, pp. 1–10, Springer, 1999. doi:10.1007/3-540-49057-4_1.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 82-91 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

It Takes Two to Tango: Merging Science and Creativity to Support Continued Innovation in the IoT Domain

Helen Hasenfuss[1, *], Muftah Fraifer[1], Sameer Kharel[1], Asma Elmangoush[2], Alan Ryan[1], Walid Elgenaidi[3]

[1]*CSIS-IDC, University of Limerick, V94 T9PX, Ireland*

[2]*The College of Industrial Technologies-CIT, Misurata, Libya*

[3]*Optical Fiber Sensors Research Center-OFSRC, University of Limerick, V94 T9PX, Ireland*

ARTICLE INFO

Article history:
Received: 28 June, 2018
Accepted: 03 September, 2018
Online: 18 September, 2018

Keywords:
IoT
Technological innovation
Infrastructure
Cloud computing
Human behavior
UX

ABSTRACT

The Internet of Things (IoT) is a new concept that has a great appeal for researchers, businesses and the ordinary tech user. It presents new possibilities of connection between devices and people and it stimulates our need and desire to interact, to exchange ideas and to communicate with the surrounding environment. This paper briefly explores aspects of the IoT that may be important for future developments and focuses on the impact of creative methodologies, such as user centered design (UCD) in a smart parking, IoT-based system prototype. It aspires to provide an alternative perspective for an improved user experience (UX) and aims to contribute to the discussion about the challenges, findings and perspectives when merging science and creativity to maintain continuous progress in the IoT domain.

1. Introduction

This paper is an extension of the article Look before you Leap presented in the conference IEEE 3rd international Forum on Research and Technologies for Society and Industry [1]. In that paper the complex system which was explored was a smart parking system through the application of the Internet of Things (IoT). It briefly discussed issues that relate to the IoT domain such as fundamental requirements, access technologies, security and privacy and data handling. The primary difference of approach in the development of this particular iteration of a smart parking system was the direct involvement of the users. The inclusion of users and/or the end users is a clearly defined methodology within design or human-computer interaction (HCI) disciplines, referred to as user centred design (UCD).

In order to preserve the quality and integrity of technological innovation it is important to combine methodologies from the science, technology, engineering and maths (STEM) disciplines and the Arts disciplines. A discussion of the methodological interplay between these diverse disciplines, as well as an attempt to expand on the challenges that emerge in the application of UCD-based methodologies into industry driven projects, are the two primary objectives of this paper. The IoT is a network of connections between digital and physical data / objects and providing a seamless user experience for IoT products is becoming an integral aspect of product design. Improved user experience (UX) is not always straightforward, as UX designers need to be knowledgeable about both the visible and invisible layers of IoT (e.g. visible layers: user interface, physical form factor and hardware spec; invisible layers: enabling architecture and technology backend, end user targets, etc.). The smart parking system that is discussed aims to highlight the impact of the involvement of the user in the design process. As IoT projects do not deal with singular issues of society (e.g. interface design, ergonomics, computing hardware, production quantity, etc.), it is necessary to shift some focus into developing multidisciplinary experts that can communicate between diverse disciplines (e.g. product design & psychology) but that are also capable of translating between academic research and industry requirements. The growth and interest in the IoT is substantial as it is estimated that in 2020, 20.8 billion devices will be connected via the Internet, according to Gartner, Inc [1]. Approximately two billion people

[*]Helen Hasenfuss, University of Limerick, Helen.Hasenfuss@ul.ie

worldwide connect to the Internet in order to avail of the variety of services ranging from playing online games, social media, commerce, education, networking sites, entertainment and many more [2].

There is great potential for the concept of the IoT because it is expanding into and incorporating a variety of diverse research areas, see Figure 1. This expansion is advantageous because of the potential advancement of peripheral technologies, e.g., developing more efficient infrastructure to improve digital and physical resource management, energy sources for nanotechnology or advancing sensor technology.

Section two, the related work, will summarise the key issues that were discussed in the related conference paper [1]. These relate to the challenges, requirements, and applications that can influence the accessibility of the IoT, especially from a UX perspective. The continuous evolution and refinement of the IoT is reflective of how technology in the computer industry developed. For example, the physical hardware of the first computers dictated the design, layout and user interaction style of the IoT. Once technology became established and grounded, researchers began to consider the user and how they interacted and engaged with the technology. A natural progression from this consideration was for the user to be directly involved in the design process itself. Since the importance and value of the user is now established, the application of UCD methodologies alongside technology design processes ensures that adaptable and sophisticated IoT products / tech can be developed without needing to apply a singular approach to design.

The user and their experience are key elements in the development of the smart parking system [1]. An analysis of why it is important to consider the user in future technological is discussed in section three. The challenges of implementing more creative methodologies such as the user centred design, agile method into industry-based projects is illustrated and the value of their contribution is explored.

In the fourth section, the smart parking system that was developed in the Interaction design centre (IDC) in the University of Limerick (UL) between 2013 and 2017 will be discussed. The approach as to how the implementation of the UCD method influenced the development of the technical and economic aspects of the project are highlighted.

The fifth section is a discussion that briefly considers elements that affect IoT based projects but also analyses whether the re-combination of STEM and Arts methodologies can make a difference in the approach to developing long term and lasting technological innovation. In an instant gratification society, endurance is becoming a rare characteristic even though it is still widely sought. The question of whether quantitative or qualitative research has the greater impact is explored from the perspective of science and artistic disciplines.

2. Related Work

It has become well known that technical challenges of IoT are diversified. In this section, we classify and identify some of these challenges.

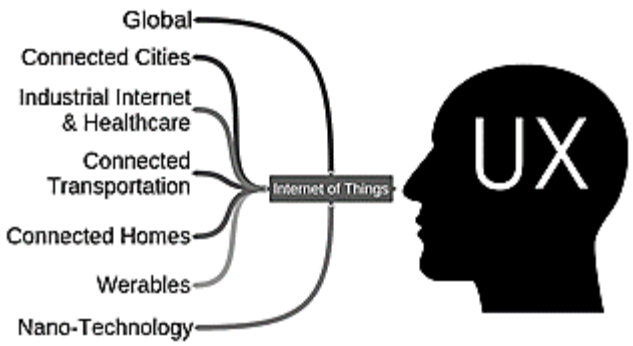


Figure 1. The influence of the IoT on local to global landscapes

2.1 Security and Privacy

One of the major concerns about IoT is security, as heterogeneous objects related to multiple domains could be connected to the Internet and communicate with various application services. Users worry that unauthorized third parties may be able to get access and/or control over their appliances and misuse data from the personal or Industrial applications [3]. Potentially, sensitive data will be roaming via the global Internet. Furthermore, most integrated components within an IoT system are characterized by low power and computation capabilities, and therefore cannot implement complex security mechanisms. IoT holds the potential of connecting everything and everyone, while not everyone should have access to everything.

2.2 Access Technologies:

Access control is necessary as well as data authentication in new communication technologies have been introduced for IoT with the objective of simplifying the deployments and gaining wider network coverage without compromising cost or power consumption. New technical solutions have been implemented that present an IP-based backhaul where the end-devices are provided with direct connectivity to a base station implementation a long-range wireless technology such as SIGFOX (SIFOX 2018) or short-range wireless technology such as Zigbee [4]. The use of the long- or short-range wireless technology depends on the application domain and the operational requirements. Table 1 summarizes the main characteristics of some other IP Protocols widely used in IoT projects. Many of these technologies are used in a variety of domains including home automation, sensor applications, and smart grids. Interested readers are referred to Naito's research for further discussion and insights of listed technologies for IoT [5].

2.3 Compatibility and longevity

The existing IoT deployments are growing in many different directions, with many different technologies competing to become the standard. Many Standard organizations have adopted a Representational State Transfer (REST) based architecture in their IoT framework such as ETSI and OneM2M, while few others have used the SOAP such as IEEE1888 [1,4]. In addition, more data protocols such as CoAP, XMPP, and MQTT are developed to address the requirements of integrating resource-constrained devices and supporting ubiquitous access.

It can potentially cause complications and might involve the deployment of additional proxies and interworking software when connecting devices.

Table 1. IoT standard and Protocols

Protocol	Max Range	Max Data Rate (Kbps)	Life of Batteries	Max Num of Devices / base station	Topology
Sigfox	50Km	1	20 years	~1Mil	star
Neul	10Km	100	10-15	n/a	star
LoRaWAN	15Km	37.5	10	~10K	Tree
Z-Wave	100m	100	Few years	232	Mesh
Zigbee	70m	20/40/250	Few years	2-6K	star
Bluetooth low-Power (BLE)	100m	1M	10-15	7	star
6LoWPAN	200m	250K	Few months	100	Mesh

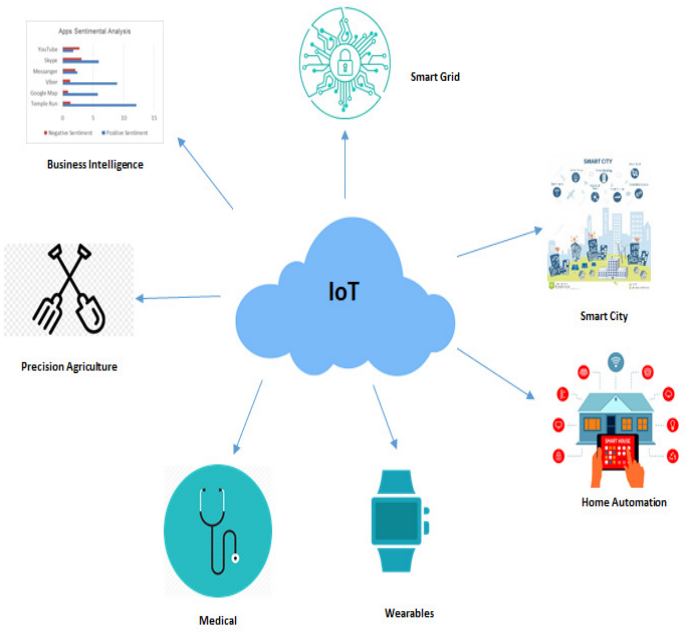


Figure 2. IoT fields

2.4 Fundamental requirement for IoT

IoT may also be a profitable venture for users in various businesses. It could potentially enable systems and smart manufacturing by connecting machines so that an industry can create networks along the entire life-time chain of a product or service [1, 6]. The Telecommunication Sector of the International Telecommunication Union (ITU-T) has specified high-level requirements of IoT in Y.2060 recommendation (ITU-T 2012).

2.5 Applications

IoT applications are now starting to integrate various domains such as personal, home, enterprise, utilities and environment, as illustrated by Figure 2.

Many services and applications have been developed, for example, in the medical domain with ubiquitous health care using sensors and body area networks to upload medical data to servers. Such home monitoring systems enable family members or community members (e.g. neighbourhood groups) to be notified about the condition of a person. This can potentially reduce the cost of health-care through pre-emptive action which can lead to an improvement in community awareness and care [7].

3. User experience challenges

IoT is growing every day and focus has been on technical aspects such as IoT infrastructure requirement, operating systems, security, power consumption, data, and storage. These technical aspects are relevant factors from an industrial perspective. However, involvement of users in any stage of development process has a positive impact on any product and improve the UX of the product. In an IoT ecosystem, different devices communicate with each other, possibly with different protocols. Different protocols and interfaces of devices have added complexity for users in its use in its ecosystem [8].

The tug of war between UX and user interfaces (UI) is an additional challenge to UX designers to provide seamless user experience without compromising security (Maiman 2015) [9]. Unification of interfaces to provide seamless experience is most challenging for UX designers in IoT. An interoperable design that can be easily adapted to the complexities of an IoT ecosystem is crucial to overcome this problem. Unification of interfaces across different appliances seems to be the biggest challenge for IoT user experience designers, as IoT products consist of different visible layers over a number of invisible layers, and designing a uniform experience requires much consideration (Janaway 2016, Staff 2016). While IoT UX can also be improved once IoT products are in market by getting feedback from users, changing physical design may not be straightforward due to its cost [9]. UX designers need to understand the different IoT invisible modules/layers that play a part in user interaction, like implementation of hardware, connectivity, data as IoT products are built on top of an IoT platform. Another challenge in interface design is that brands can have different requirements for the same software when used on different devices for example Android and iOS devices. Designing a UI for the IoT can be challenging since it, in many cases, is highly interconnected with other products, systems and services which affects the users' perceptions of their experiences.

The IoT can have a positive impact in the life of many citizens (users). It can be used to study human behaviour and understand how technologies are used in a daily life and this study can be useful for further IoT technology design. Involvement of the user via user centred design (UCD) in IoT design can't be ignored. The majority of IoT projects are still based in a lab setting, however ideally testing should be performed in real scenarios with real end users to find real problems [10].

3.1 Activity Centred Design (ACD) vs User Centred Design (UCD)

It is argued that the concept of UCD is not relevant in case of IoT. Cruickshank, L. and N. Trivedi suggest that the human should be seen as one of the smart objects in the IoT ecosystem and should fit into an activity centered design (ACD) approach. Cruickshank, L. and N. Trivedi argues that overall, the contribution of non-human elements is significantly higher than purely human elements in IoT systems. The discussion in academia relates to whether design should be able to keep up with emerging new agency rather than just try to satisfy users/customers [11].

In considering the ACD concept, we shouldn't forget that we are not trying to build robotic world but create robots for facilitating their (user/customer) daily life, maintaining privacy and without having tension between device and human. The results from Memedi's, research clearly shows that UCD plays important role in IoT. They found UCD as a helpful tool for identifying the requirement of the system user. For example, in a system developed for Parkinson's disease, patients it was found that 70% of users reacted positively to a IoT-based system prototype, (connected devices such as a wrist sensor, smartphone, bed sensor and an electronic dosing device). Positive impact of the system is due to including user in application/system development [12]. It was found that challenge of communication has been one of major component for hindering implementation of UX in the IoT domain. It was found that users are concerned about communication between device invisibly and users are unaware between information shared between it and different objects. It was also found users are interested in whether value is added to their life by connecting devices. Sometimes the complexity of IoT device configurations can also effect UX [13].

There are a variety of aspects that influence consumer acceptance of IoT systems and products. Factors that directly affect UX, range from tangible to nonphysical issues. For example, device memory and power: using many different IoT apps can be a drain on mobile devices or if it takes longer to complete a task using an IoT app than it would if it were done manually. Errors in IoT products can be detrimental to user experience, because the user can become irritated and generate more negative affect (see subsection 5.4). As a result, the user could simply switch to a similar product provided by an alternative company. Alternatively, qualities such as usefulness and trust have varying degrees of priority in relation to acceptance of IoT technology but are primarily dependant on the interaction and perception of the user.

Poor user experience for many IoT products is often due to ineffectual user-testing during the design process, leading to badly designed interfaces of applications and devices. For example, the smart heating system designed for supporting energy-saving lacked in all metrics used for its UX evaluation [14]. Users experienced difficulty accomplishing their tasks and the system also had inappropriate labelling [14]. Current trend of software development method has been developing rapidly, iteratively and incrementally and delivering modules of products to customer in short timeframes. The process is popularly known as "Agile".

3.2 How practical is Agile in conjunction with UCD?

The IDC in the University of Limerick, conducted interviews with 10 IT professionals working in software companies as UX experts, designers and software developers from Ireland, Nepal and United State of America from August to October 2017. Every interviewee agreed that in some way they were following the Agile method and among different Agile methods "Scrum" is used in most cases. Agile is an umbrella term used for iterative and increment software development methodologies. Scrum is one of the sub-methods under Agile. In Scrum in every two to four weeks one complete module of the product is delivered to the customer which is known as a sprint.

From the above paragraphs it is clear that UCD should ideally be a mandatory concept in the development of user friendly and seamless UX in IoT devices. Brhel, M., et al., Butt, S. M. and S. M. Butt, Sohaib, O. and K. Khan have been writing about integrating the concept of UCD into the Agile development method [15-17]. A fundamental difference between agile development method and UCD is that the former focuses on the customer and the latter on the user. User and customer are almost the same, but every customer may not be an end user. For example, from a business perspective, everyone is a customer but, in our case, "user" refers to specific or targeted end users of a particular product, i.e. for a smart parking system based in UL – drivers who knew the campus grounds and administration were the appropriate users. To improve UX, suggestions have been made to use the Agile methodology to develop IoT devices in conjunction with UCD [18].

Regarding how often the software industry actually uses the concept of UCD in the development of new products, from the previously mentioned interviews of 10 IT specialists, we learned that incorporating UCD is not a straightforward process. The IT specialists agreed that implementation of UCD depends upon,

A) budget of the project,
B) requirement of customers and
C) that it is the responsible of the customer to know about their users. (In this case customer refers to an organization who enlists the services of a company that specialises in developing IoT products and systems.)

In [19], the paper proposes that usability testing is the least preferable testing method compared with unit, integration, system and acceptance testing as this testing are done by technical experts rather than involving end users. Mostly usability testing is conducted by software engineers rather than usability or UX experts. They indicate that customers give less priority to usability testing rather than acceptance testing, alpha and beta testing. Lack of usability is often attributed to the following aspects: 35% due to lack of time, 20% due to lack of training/knowledge, 15% lack of budget and rest for other reasons [19]. It state that *"companies generally do not give priority to usability and UX evaluation"* and including the user in software development is taken as a burden or delay in time and increased cost for relatively small benefit [19]. Different UCD tools have be implied in the Agile methodology for usability and UX. The usability tools vary slightly according to the discipline in which they are applied. For example, in requirements engineering, ethnographical observation, card sorting, personas, task scenarios, scenario and storyboards, and prototyping are the primary tools used to evaluate usability. In the design discipline: screen snapshots, product style guide, navigation maps are used. In the evaluation process, heuristic evaluation, cognitive

walkthrough, thinking aloud, and lab usability testing are implemented. Then, an evaluation by experts, well-designed usability tests and analysis of installed systems are processes used as defined by UCD or have been slightly modified according to project demand [20].

The process of developing a smart parking system by including the relevant users or stakeholder and the positive outcome of the project highlight that participatory- and user centred design methodologies are invaluable for the overall design process but in particular the early stages of project development. This postulation is supported through diverse academic literature [10, 12]. Consistency, intuitive design and good, logical mapping are just some of the cornerstones for the successful uptake of any design as illustrated by Don Norman in his book "The Design of Everyday Things" [21]. Both lab and field testing need to be performed on IoT devices with users. Lab testing is helpful to improve the design of products and field testing is helpful to improve the user experience. Different user experience evaluation techniques, such as contextual laddering, day reconstruction, experience sampling and UX curves, can be applied as per requirements.

3.3 Discussion of identified research gaps when applying UCD

One of the basic steps in using the UCD method is to know everything related to the existing problems of the system expected to be designed. In order to build up this knowledge base, initial research should be carried out on other systems that have the same purpose. Whilst research indicates that there is a need for smart parking system solutions, there were no previous studies published in the literature relating to the assessment of these systems that also take into account the views of users. As well as this research indicated that each solution had different technological requirements depending on which factors received the highest priority, e.g. cost effective, integration possibilities, interface, etc. Each prototype system has advantages and disadvantages in terms of the following criteria:

- cost
- reliability
- scalability
- accuracy
- communication type
- circuit complexity
- reliability
- method of operation
- ease of installation
- usability of the system

There is an inadequate research base to fully inform how to design and evaluate smart parking systems for users.

4. Challenges and findings from our perspective in the IoT domain

UCD is a broad term to describe a design process based on the integration of potential users and since this era is witnessing a huge technological revolution of future networks depends on the connecting different objects from watches, cars, to satellites. There is no question about it: technology has revolutionized every part of our life nowadays as the Internet is likely to have a dramatic impact on our daily lives as it evolves into an essential part of many systems such as smart transportation, and smart parking etc.

The dialectical question that exists among those interested in the IoT domain remains: do we really need to link all things to each other? Because the different answers to the same question have advantages and disadvantages to the benefits of that linkage. However, the stakeholder is the only one to be able to estimate the extent to which he can benefit from the possibility of linking different things together to achieve the desired goals. Enabling technology from the things around us to be able to interact with each other and with the user, has many advantages and disadvantages that require careful consideration before any system is designed to deliver its services in a meaningful manner. We see an urgent need to have a harmony between merging science and creativity to support continued innovation in the IoT domain.
Certainly, exploiting the things around them and looking for alternative solutions to improve the level of services that users need is a great challenge, especially with regard to longevity and sustainability. This will create solutions for many dilemmas and problems that people face in general in various aspects of life.

4.1 IoT based smart parking system

Fraifer et al., designed a smart system that helps to find parking spaces at the University of Limerick. The difference here is that they incorporated the user into the design process in an early process of user-centered design [22]. Three groups of users were used to explore the problem of finding parking on an enclosed but spread-out area and to design a series prototypes to resolve, or alleviate, the issues. These issues include,

- reducing pollution by reducing the amount of time drivers spend looking for parking spaces,
- finding spaces hidden by larger cars,
- real-time update of free parking spaces,
- discerning peak times for searching, parking and when spaces are freed up, etc.

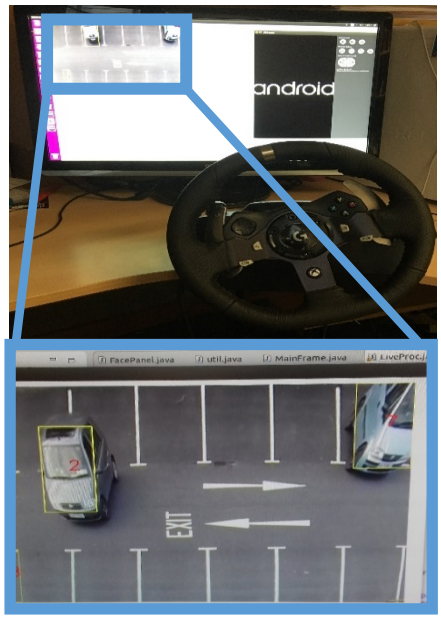


Figure 3. Illustrates how the CCTV footage was integrated in order create a search algorithm (left) and smart parking app (right of screen).

The basis of the design implements CCTV which is often installed in any parking lot for the purpose of securing and preventing theft making it a viable economic solution. The user interacts with the system through a specific application in which they can see which places are vacant and where they can book a parking space [22], see Figure 3. The study has presented a design development of an application of IoT technologies. This has provided some findings that challenges the traditional approach to designing IoT based systems, e.g. how to make abstract technologies connective and interactable. Such findings helped to identify key areas of future research and intervention. They also support the argument for the need of new ideas and solutions that are mindful of the driver's requirements and how to improve the potential IoT smart parking system. Below is analysis data and finding discussion from two iterations of UCD approach.

4.2 UCD methodology

The UCD methodology frames the phases throughout a design and development of the prototype on gaining a deep understanding of those who will be using the system.

- **Assigning the context of use:** 12 members of the UL community participated in the smart parking project. These users were on campus on a daily basis and regularly experienced issues in finding parking. They include students, staff and frequent visitors
- **Specify requirements**: The users were interviewed and asked to complete a questionnaire regarding the primary issues they experienced in their normal parking routine. Within this phase it is possible to discern the major issues that should be addressed throughout the project but also what the users envisioned could potentially alleviate these issues.
- **Create design solutions:** Within this phase the majority of design work takes places and the actual product comes to fruition. For the smart parking project an app was developed that would enable users to find and reserve parking spaces cutting down on the time and fuel spent cruising around the large campus in order to find a parking space.
- **Evaluate designs:** This last phase is the crucial difference a UCD approach can make to product development. The evaluation procedure can be applied often throughout the design process and does not always require the full number of users. For example, once a low fidelity paper prototype was developed, that sufficiently conveyed the idea of the app, it was possible to evaluate it with some of the users. After this evaluation, the requirements, solution and evaluation phases can be continuously re-applied, until the final product becomes more refined and of a high-quality finish.

4.3 Findings from designing an IoT based smart parking system at IDC

The data analysis was based on qualitative findings from a human-computer interaction perspective. From an user perspective, there is an inadequate research base to inform fully how to design and evaluate smart parking systems for users; therefore, this issue requires further investigation. A summary of the results and findings of research contributors from IDC are detailed in the following subsection 4.3.1 and 4.3.2.

4.3.1 First iteration of UCD

One of the most important benefits of this phase is the discovery of errors in the initial design of the system that would have led to failure if it had been left undetected. In addition, using different categories has main important benefits that has had beneficial results to redesign the prototype according to recommendations from users. The aim was to know the basic dilemma facing drivers while searching for a place to park. Other ideas relate to the development of solutions and ideas that may be part of the system.

4.3.2 Second iteration of UCD

The second iteration was used by a small group (focus group) consisting of 3 users. They were selected to perform a second session of UCD regarding the improvement of the system after the previous proposed modifications. This stage leads to the emergence of proposed system as a supposed final product. One of the findings in this stage and through the discussion with users, is that the system must include several things that contribute to solving the dilemma of finding a parking space. Below are the findings based on the data collection when applying the UCD approach. The initial model was an experiment in an indoor environment. The car park was designed using cardboard, where the camera was set up to face the simulated parking lot.

- The first idea was to draw circles with dark colours on every parking lot in order to facilitate the process of identifying whether the position is vacant or not. The system depended on computer vision, and it is easy to identify shapes by using the functions available in the OpenCV library.
- After presenting the model to the users, some of them made the comment that the system may not work in the event of snowfall as the circular shape is covered in winter by snow or autumn leaves. Users came up with the idea using actual cars, as that would be analogous to rectangles, using computer vision to detect if the parking lot is vacant.
- With regard to the designed mobile application, which is a fundamental element of the system, think-aloud sessions with users and two iterations of UCD enriched the appearance and usability of the model.
- The general tendency of participants was to make the application simple, but 9 users believed that the application would be more effective if it would show parking without activating reservation. They believed that using the application while driving may pose a risk to the driver. Enabling the driver to know the available vacant positions is in itself a good solution.
- To navigate directly to the vacant space to save time, fuel and mental effort, which is a basic goal of the system.

The primary concern of the 9 participants was that the system does not contain barriers to guarantee a parking space and gave the impression that the general system may take some time to function properly. Some users suggested activating the possibility of booking a specific space. In the cases where there is a fee for booking the space, the administration would be responsible for the space, should someone take it without a reservation. Developing scenarios often identifies important aspects of using a product in the real world that may not otherwise identified and considered in a lab setting, for example, the workflow being disrupted by a phone

call. Scenarios are useful throughout the design process, particularly when developing task descriptions for usability testing.

4.4 Findings from designing an IoT based smart parking system at IDC

The primary findings from the smart parking system at the IDC relate to the infrastructure which the system uses and how the users felt about using the system. The infrastructures (Wi-Fi and CCTV cameras) which this system used enabled a smooth integration and reliable service. Internet access is widespread throughout the campus and CCTV is used for campus security. In general, the users were happy with the system and the app that was developed because it was easy and quick to use and did not impede their searching process. Advantages of this system are the ability to find available parking spots, especially when they are not visible from a driving position, and that the number and location of free spaces are updated in real time. Table 2 summarizes the most important aspects that were defined by users in the 2nd round of evaluation [27].

Table 2. Findings from 2nd iteration of evaluating the smart parking system.

Source	Needs/Requirements/Features/ Solutions/ /Ideations	Findings
Student	Using Google Maps API library services to determine the destination accurately, especially in large parking areas. The possibility of seeing the position of the car in a real way by connecting the system to live broadcast, perhaps through live images.	Providing a precise positioning service such as Google Maps will be better for the user (user - 01 Second session) The availability of a service enables the user to see parking spaces in real form in the form of images that have a positive impact on the user when he/she arrives (user - 01)
Visitor	Create a database for users by asking the same user to enter his/her favourite times to take advantage of the vacant places for others if possible.	The availability of database users will help administrators to provide available vacant spaces smoothly, (user - 02)
Staff	The possibility of sending short messages or notifications when a user leaves a place before the end of the reservation. Provide information to the user about each position on the application, for example a counter in descending order to make the rest of the users aware of the nearest position of silent vacant	Enable the users to add the SMS of informing others that user leaves the pre-booked place before the end of the reservation. (user - 03) The availability of a digital counter in descending order showing the end of the booking represents a beautiful feature to remind the user, (user - 03)

5. Discussion

One challenge of technological innovation is its uptake into the public domain. The flexibility and adaptability inherent in the IoT concepts presents a myriad of new opportunities to find eloquent solutions as technology becomes more sophisticated. However, user integration into the technology design process is still a major obstacle. After a company identifies the appropriate target user group, which in itself can be challenging, extracting the relevant user data can be a deterring factor for the majority of companies. Collecting user data is often time and resource consuming and it is necessary to translate the user's requirements and opinions into data that a company can use. As has been highlighted throughout this paper it is necessary to,

A) Engage the users from early on in the design and development process for any IoT potential products or services,
B) Adapt the existing business models to support cost-effective services, [1]

In addition to adapting existing models in industry, it is also necessary to consider the methodologies used in academic research. Whilst this concept is the primary focus in the discussion section, subsections 5.1 to 5.4 briefly summarise previously addressed issues [1] in order to provide context for the extended discussion on the merging of science and creativity to continuing progressing technological innovation. These issues are as follows:

- Business requirements
- Existing infrastructure
- Global connectivity and networking
- Emotion-driven design

5.1 Business requirements

The IoT refers to uniquely identifiable objects and their virtual representations in an Internet-like structure. Special requirements are needed in order to manage, secure and facilitate these embedded software assets. These systems are considered an industrial subset of the IoT, which may include billions of interconnected sensors, devices and systems, many of which will communicate without human involvement [1].

5.2 Existing infrastructure

The global and expanding nature of the IoT means that the viability of the network in its entirety is an important factor. For such a dynamic system to continuing growing and expanding it is necessary to consider how this viability can be maintained and is affected by the diverse number of variables of which it is comprised, i.e. the system's stability, the existing infrastructure and availability of resources (bandwidth, storage capacity, IP address space, etc.). Similar to any human endeavour having a good grounding or baseline is important, e.g. grammar enables a person to gain command of a language and to understand it. In relation to the IoT, since it is built upon and using the existing infrastructure of the Internet it is necessary to keep reviewing the efficiency of the technologies involved. For example, if an older infrastructure can no longer support new technologies because either conceptual or physical compatibility between technologies is no longer possible, too many resources may be required in order to maintain it. This action would shift the focus from development, i.e. looking forward, to maintenance, i.e. being tied to the present and past. It correlates with Mark Weiser's description that a "good tool is an invisible tool." [23]. For example, becoming conscious of a failing infrastructure (glitches, connectivity issues, slow connection, limited development opportunities, etc) detracts from an actual IoT device or system. A good tool ensures that the user's focus is entirely on the task.

5.3 Global Connectivity

The IoT is a network of connected devices and a key issue of connectivity is compatibility. To accommodate the question of compatibility raises questions regarding the concept of standardisation. In a system that is inherently built to accommodate diversity, this approach appears to be contradictory. Even though standards can ease the element of interaction with technology (e.g. the classic floppy disk as the save icon), it is necessary to maintain a balance between system diversity and strength. For example, allowing systems and devices to be compatible with each other can create an IoT network that is stable in the event of infrastructure degradation or failure due to physical factors, (e.g. power outage, out of range, no connection, no storage, etc.) [24]. Allowing the method of connectivity to be diverse can ensure that the whole system is stronger against attack and more adaptable. The latter option is also more analogous to the method in which groups of people interact with each other, e.g. language dialects.

Considering the IoT development from human aspects, such as communication, growth, and development, the interdependent nature of this type of network or system becomes evident. The IoT aims to be integrated into an environment that is inhabited by social, living entities. As a result, logic and reason must interact with emotion and unpredictability, e.g. technical developments are directly influenced by user experiences and vice versa. The social and communicative aspects that are inherent in the IoT system creates spaces for large as well as small communities to work together in order to generate smaller networks that can eventually be integrated or expanded further. People can gain a sense of "socio-technical responsibility" (de Rosney 2014) and can break down knowledge barriers.

5.4 Emotional design

The social aspect mentioned in section 5.3 highlights an intriguing challenge particularly from a design perspective: the unpredictability of human nature. The ideas put forward in phenomenology illustrate how multifaceted a single experience can be. Aside from design considerations such as ergonomics, emotion and thereby experience generation and perception are critical elements in the design process. In the early years of computer development, technology defined user interaction styles. Nowadays the awareness of the importance of designing for an experience as well as for efficient technology is the progress of technological development that is embodied in the IoT. Previously a user had to adapt their behaviour according to the technology. As the focus in industry shifted from pure technology development to aesthetics and user experience it was possible to see an adaptation in the design process. Instead of primarily emphasising the engineering and scientific methodologies, it became necessary to consider socio- and design-based methodologies. A part of the user experience is the emotion involved in the interaction with technology [25]. Automated teller machines (ATMs) are often referenced as a user interface whereby clear mapping, ease of use, informative feedback and an aesthetic yet practical design are essential qualities. Studies illustrate that an interface embodying these characteristics ensures that users interact with more patience and tolerance and that they have a greater capacity for creative problem solving should an error occur. Overly-complicated and frustrating designs can evoke "negative affect". As a result, users tend not to engage long enough in order to find appropriate solutions thereby creating negative experiences [26]. It is important to note that negative affect can provide valuable insights regarding developing interfaces that may also need to be used in times of emergency, whereby stress, pressure, or panic can seriously hinder decision-making processes. Considering a user's emotions and experiences in the design process, provides a greater understanding of the effects of human emotional states which in turn can aid in the development of technology that behaves more intuitively.

5.5 Closing the gap between logic and creativity

Identifying the appropriate stake holders is a key element once a UCD approach is implemented. Targeting the right users will be most cost effective and provide valuable and accurate data. For example, testing a sophisticated gaming interface (e.g. joystick) on users who don't play computer games will not yield accurate useable data. Not engaging or interacting with the target group of end-users throughout the design process could be detrimental because it could result in a design or product that does not meet user requirements or is not commercially viable. A key component to UCD, aside from the involvement of the user, is the continuous evaluation process with the user, of the design itself. The repeated evaluation ensures that the primary aims, and requirements of the user are always up to date.

Design based methodologies are steps towards bridging the gap between scientific procedures and practice based artistic explorations. Merging science and creativity to support continued innovation in the IoT domain reflects the incorporation of new ideas that emerge through research into practical applications in the STEM disciplines. With respect to long term innovation, an argument can be made for a more active application of the STEAM approach as opposed to a purely STEM orientated procedure (the A in STEAM representing Arts). The generation of new ideas in research reflects in part the creative element of the human psyche [26].

The inclusion of the procedures evident in an artistic methodology has the potential to enrich those of the more logic and reason-based disciplines. It is evident in history that there was little differentiation made between the arts and sciences, however an intuitive method of discovering or validating certain truths is based primarily on deductive reasoning. This type of reasoning involves determining specific facts and narrowing the field of exploration, i.e. number of variables that affect the system, in order to ascertain specific and conclusive truths. An analogy can be drawn to top-down processing: a researcher may start with several questions but eventually ends with one singular focus. The STEM disciplines mainly embody this type of reasoning approach and has been invaluable in defining specific procedures that facilitate repeatability, consistency and transparency in good scientific research [26].

An issue with an approach that encourages an exclusive focus on singular elements of a system or product, is that the research scope can become stale, it can miss peripheral yet related subjects and can become too rigid and unforgiving. For these reasons it is important to consider the methodologies applied in the arts, which

in contrast to scientific or engineering disciplines are primarily based on inductive reasoning. Inductive reasoning embodies a bottom-up approach to research, i.e. it is the process whereby knowledge is acquired and constructed in order to generate questions that are further explored in the deductive process. An output of inductive inference are the probabilities of certain truths. For example, conclusions can be drawn logically from the knowledge acquired but it does not necessarily make accommodations if the knowledge base is inaccurate or incorrect. Therefore, the truths that emerge from inductive inference are a source of ideas and inspiration that can enrich the disciplines in which deductive reasoning is favoured [26].

The arts are primarily based on creativity - the ability to work with loose boundary conditions, the unknown or the unpredictable, to be able to see connections between diverse subject matter and to view errors as opportunities and not always as dead-ends. Similar to the sciences, a pure focus within the arts, e.g. dance, poetry, drawing, enabled discipline-specific methodologies to emerge (e.g. social art, practice-based methodologies). However, it is important to recognise that these methodologies in isolation are as self-limiting as those in the STEM disciplines.

The difference between logic and creative based disciplines is also reflected in the form of quantitative and qualitative output. Independently, each type of output has its value, but they also represent the different perspectives through which data can be interpreted. Therefore, it is not necessarily possible to declare that one form of output is better than the other, just as there is no one main research methodology that stands above others, but it is important for researchers to remain aware of these perspectives when using data and to apply a degree of objectivity.

In relation to long term technological innovation, a degree of time and energy must be invested into the early stages of technological development. In the generation of questions, it is important to draw conclusions that support a high probability of truth before the deductive approach is applied. Achieving this state indicates that the research must be thorough and accurate. The following example highlights the number of steps involved in taking an idea through the UCD process to create a high-fidelity prototype that is ready for industrial or public implementation. The even distribution of energy throughout these processes ensures that the end result is of good, reliable and repeatable quality.

- Idea generation through brainstorming with design team and ideally with relevant users
- Sorting of ideas into relevancy, possible pitfalls, new ideas, do's and don'ts
- Research one or two specific ideas and translate from idea into tangible concept
- Create a low fidelity prototype
- Run another brainstorming session with the prototype or run individual interviews
- Re-sort new contributions and research design and product elements further (e.g. material, platforms, hardware, etc.)
- Adjust and adapt the prototype to incorporate the feedback
- Run another evaluation session with the relevant users
- After the comments and feedback have been incorporated again it is possible to develop the prototype further.
- Create a high-fidelity prototype
- Implement the prototype and get real world data on interaction, glitches, successes, improvements, etc.
- Refine the prototype accordingly
- Run final tests with the prototype and evaluate the system via interviews with users
- At this point the prototype is thoroughly tested and can be moved on to the industrial phase whereby it could potentially be implemented beyond the test area.

The user centered design approach is one method by which an even distribution of energy and resources can be implemented into industry-based projects. Even though this process can initially take more time and resources, the results usually yield more accurate market research, public uptake and better quality of outputs, i.e. short-term versus long-term gain. The smart parking project described in this paper is an example of how important it is to consider the logical and creative aspects of the human psyche in the future development of technology.

The merging of logic and creativity advocates a flexible and adaptable approach to research. Whilst it is necessary to continue having singular-subject experts, to support technological innovation it is also necessary to train multidisciplinary experts. The latter type of expert and / or researcher develops the skill to communicate and translate between two or more diverse disciplines (e.g. biology, 3D printing and modelling). An advantage of this skill is the ability to alter perspective. It allows researchers to make connections between subjects or problems, etc, more quickly and they can bridge the gap between deductive and inductive inference.

6. Conclusion

The application and integration of design methodologies is a step towards integrating creative processes into scientific disciplines. Drawing out the skills in order to generate new ideas is vital in the enrichment of research scope. It is clear that the process of inductive inference (idea generation) and deductive inference (translating ideas into reality) is continuous and circular, as one phase feeds into the other. For this cyclical process to be successful it is necessary to find a balance in the application of creative and logical methodologies. In contrast focusing on either extreme (i.e. purely logic-based or creative-based methodologies) enables researchers to explore boundaries and limits of a specific discipline [26]. The peripheral topics that have been mentioned in this paper (e.g. business objectives, security, privacy, connection, etc) are not a conclusive list but demonstrate the range of interdependencies in a domain as versatile as the IoT. Even though the topics may not be directly involved, their influence can have interesting effects and outcomes on both the system in question and the related users. It is a constant interplay between variables that are known to the designer and variables that emerge as a result of implementation and real-life interaction. A chance always exists that designs may be used, not for what they were originally intended, but possibly for more creative or destructive purposes. It is another compelling reason for including users in

the product development process as early as possible because even though it may initially cost more with respect to resources and time, the long-term benefits are evident in the quality and efficiency of products produced.

The smart parking system detailed in this paper explores the relation between the involvement of the user and the development of the technology. The results from this project are congruent with research studies on system development and usability, to facilitate user acceptance and efficiency, accuracy, and satisfaction. Because the users were involved through all four phases of the design process, their needs, suggestions, and preferences were continuously incorporated into the design and evaluation of an IoT based smart parking prototype. The UCD methodology transforms users into active contributors to the design process and has provided improved UX in each iteration. It enables the final product to be tailored directly around users' needs rather than providing superfluous functions that disrupt the design.

Acknowledgments

We would like to acknowledge the support of the Ministry of High Education Scientific Research-Tripoli, Erasmus Mundus – LEADERS and Interaction Design Centre (IDC) in CSIS department at University of Limerick for their funding of the project.

References

[1] M. Fraifer, S. Kharel, H. Hasenfuss, A. Elmangoush, A. Ryan, W. Elgenaidi, and M. Fernström, "Look before you leap: exploring the challenges of technology and user experience in the internet of things" In Research and Technologies for Society and Industry (RTSI), 2017 IEEE 3rd International Forum on, 1-6, IEEE, Modena, Italy, 2017. https://doi.org/ 10.1109/RTSI.2017.8065920

[2] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges" Ad hoc networks, 10(7), 1497-1516, 2012, https://doi.org/10.1016/j.adhoc.2012.02.016

[3] A. Grau, "Can you trust your fridge?" IEEE Spectrum, 52(3), 50-56, 2015. https://doi.org/10.1109/MSPEC.2015.7049440

[4] M. Tariq, Z. Zhou, J. Wu, M. Macuha, and T. Sato, "Smart grid standards for home and building automation" in 2012 IEEE International Conference on Power System Technology (POWERCON), pp. 1–6, Auckland, New Zealand, 2012. https://doi.org/10.1109/PowerCon.2012.6401448

[5] K. Naito, "A Survey on the Internet-of-Things: Standards, Challenges and Future Prospects" J. Inf. Process., vol. 25, pp. 23–31, 2017. https://doi.org/10.2197/ipsjjip.25.23

[6] D. Uckelmann, M. Harrison, F. Michahelles, "An Architectural Approach Towards the Future Internet of Things" In: Uckelmann D., Harrison M., Michahelles F. (eds) Architecting the Internet of Things. Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-19157-2_1

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions" Future generation computer systems, 29(7), 1645-1660, 2013 https://doi.org/10.1016/j.future.2013.01.010

[8] M. Dixit, J. Kumar, and R. Kumar, "Internet of things and its challenges" in 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 810-814, IEEE, Noida, India, 2015.https://doi.org/10.1109/ICGCIoT.2015.7380574.

[9] J. Bergman, T. Olsson, I. Johansson, and K. Rassmus-Gröhn, "An exploratory study on how Internet of Things developing companies handle User Experience Requirements" in International Working Conference on Requirements Engineering: Foundation for Software Quality, 20-36, Springer, Cham, Utrecht, The Netherlands, 2018. https://doi.org/10.1007/978-3-319-77243-1_2

[10] M. Nati, A. Gluhak, H. Abangar, and W. Headley, "Smartcampus: A user-centric testbed for internet of things experimentation" In Wireless Personal Multimedia Communications (WPMC), 16th International Symposium on, 1-6, IEEE, Atlantic City, NJ, USA 2013. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6618632

[11] L. Cruickshank, and N. Trivedi, "Beyond Human-Centred Design: Supporting a New Materiality in the Internet of Things, or How to Design When a Toaster is One of Your Users" The Design Journal, 20(5), 561-576, 2017. https://doi.org/10.1080/14606925.2017.1349381

[12] M. Memedi, G. Tshering, M. Fogelberg, I. Jusufi, E. Kolkowska, and G. Klein, "An Interface for IoT: Feeding Back Health-Related Data to Parkinson's Disease Patients" Journal of Sensor and Actuator Networks, 7(1),14, 2018 https://doi.org/10.3390/jsan7010014

[13] J. Bergman, and I. Johansson, "The user experience perspective of Internet of Things development," Master thesis, Lund University, 2017.

[14] S. Wall, and F. Healy, Usability testing of smarter heating controls. A report to the Department for Energy and Climate Change. Amberlight. DECC, London, 2013https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/266220/usability_testing_smarter_heating_controls.pdf

[15] M. Brhel, H. Meth, A. Maedche, and K. Werder, "Exploring principles of user-centered agile software development: A literature review" Information and Software Technology, 61, 163-181, 2015. https://doi.org/10.1016/j.infsof.2015.01.004

[16] S.M. Butt, and S.M. Butt, "Usability Evaluation Method for Agile Software Development" International Journal of Software Engineering and Computer Systems, 1(1), 29-40, 2015. http://dx.doi.org/10.15282/ijsecs.1.2015.3.0003

[17] O. Sohaib, and K. Khan, "Integrating usability engineering and agile software development: A literature review" In Computer design and applications (ICCDA), 2010 international conference on (Vol. 2, pp. V2-32). IEEE, Qinhuangdao, China, 2010. https://doi.org/10.1109/ICCDA.2010.5540916

[18] X. Larrucea, A. Combelles, J. Favaro, and K. Taneja, K., "Software engineering for the internet of things" IEEE Software, (1), 24-28, 2017. https://doi.org/10.1109/MS.2017.28

[19] M.K Larusdottir, E.R. Bjarnadottir , J. Gulliksen, "The Focus on Usability in Testing Practices in Industry. In: Forbrig P., Paternó F., Mark Pejtersen A." Human-Computer Interaction, vol 332. Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-15231-3_11

[20] D.A.Magües, J.W. Castro, and S.T. Acuna, "HCI usability techniques in agile development" In Automatica (ICA-ACCA), IEEE International Conference,pp. 1-7, IEEE, Curico, Chile, 2016. https://doi.org/ 10.1109/ICA-ACCA.2016.7778513

[21] D. Norman, The design of everyday things: Revised and expanded edition, Constellation, 2013.

[22] M. Fraifer , H. Hasenfuss, M. Fernström, "Taking Away the Green Screen—A Brief Discussion of a Multidisciplinary Approach to IoT via, Smart Parking and Interaction Design" In Smart Trends in Systems, Security and Sustainability, 33-45, Springer, Singapore, 2018. https://doi.org/10.1007/978-981-10-6916-1_4

[23] M.Weiser, "The world is not a desktop", interactions, 1(1), 7-8, 1994. https://doi.org/ 10.1145/174800.174801

[24] J.O Kephart, and.M. Chess, "The vision of autonomic computing" Computer, (1), 41-50, 2003. https://doi.org/ 10.1109/MC.2003.1160055

[25] D. A. Norman, Emotional Design: Why We Love (or Hate) Everyday Things, Basic Civitas Books, 2004.

[26] H. Hasenfuss, "A design exploration of an agent template for multiagent systems (MAS) for shape shifting tangible user interfaces," PhD Thesis, University of Limerick, 2018.

[27] M. Fraifer, "A design exploration of an IoT based smart parking system with stakeholders using User-Centred Design" unpublished PhD Thesis, University of Limerick, 2018

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 83-88 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Enhanced Ship Energy Efficiency by Using Marine Box Coolers

Abdallah Aijjou*, Lhoussain Bahatti, Abdelhadi Raihani

Laboratory: signals, distributed systems and Artificial Intelligence, ENSET Mohammedia, University Hassan II Morocco.

ARTICLE INFO

Article history:
Received: 13 August, 2018
Accepted: 24 October, 2018
Online: 01 November, 2018

Keywords:
Ship energy efficiency
Box / keel coolers
Central cooling
Sea water pumping

ABSTRACT

Climate change, increasing fuel oil prices and new international regulation on ship emissions lead to more focus on shipping fuel consumption and energy efficiency. There are various solutions for improving the ship energy efficiency. In this manuscript, we aim to present a real case of energy saving by adopting the central cooling system with box cooler on the ship instead of conventional system. The electric energy power necessary for operating the machinery cooling system of the ship is calculated for conventional cooling system and compared to the cooling system using box coolers in term of fuel oil consumption and CO2 emissions. This study quantified the fuel saving potential that could be achieved with use of keel coolers. Adopting central cooling with box coolers may contribute in reduction of fuel oil consumption and improving the ship energy efficiency. systems. *should not contain citations.*

1. Introduction

The ship machinery is essentially composed by propulsion diesel engine, auxiliary engines for electric power production and boiler for heating purposes. Other auxiliary equipment necessary for the operation are also fitted such air compressors, air conditioning plant, steam condensers, hydraulic power pack for ship mooring and cargo operation etc.

All this equipment produces undesirable heat and need cooling which remove excessive heat out of surfaces of material, safeguard the metal mechanical properties and keep the temperatures within the limits specified by makers for maximum performances during operations.

The propulsion plant (mostly diesel engines) is the largest source of energy to dissipates by cooling system [1]. The cooling system is designed to covers machinery cooling needs when most of the equipment are in operation at its maximum powers. Safe margin is also added.

When we look at the electric energy balance of various types of ships we notice that cooling water pumps are among the largest and the most equipment running over the time, since they are solicited either when the vessel is at sea or alongside. Therefore, cooling pumps are the largest electric power consumers on board and are accounted for 10% of the kilowatt- hour consumed.

This fact explains why several researches dedicated to ship energy efficiency focus on cooling system components. These studies have focused mainly on improving the performance and efficiency of the conventional cooling system (piping and sea water pumps) i.e. [2,3,4,5] studies on energy saving by improving the pumping system, [6] diagnosis and corrosion protection, [7] cooling system reliability.

Very few studies discussed the possibilities of elimination of sea water pipes and pumps from cooling system [8]. The use of box or keel coolers is one of the solutions available for the ship building industry.

Adopting box coolers for cooling systems allows to eliminate the pumps, filters, valves and reduces the piping length therefore the installation is more simple and cheaper. In addition, elimination of these components will reduce also the cost of the operational maintenance.

This paper aims to demonstrate the effect of box cooler use on ship energy efficiency hence after this introduction we give an overview of different ship cooling systems in section II, in section III box and keel coolers principles are described, in section IV we study an application of box cooler for real ship in operation. Conclusion is given in section V.

2. Ship machinery cooling system:

There are three basic types of cooling system commonly used in the marine machinery on board the vessel.

1-Direct cooling system with sea water (Fig.1): sea water is used as cooling media for heat exchangers in open circuit. Sea

*Abdallah Aijjou, Email: thalassa1310@yahoo.fr, Tel: +212 661 42 34 17,

water is rawen from sea and pumed directly through the machinery system before being discharged overboard.

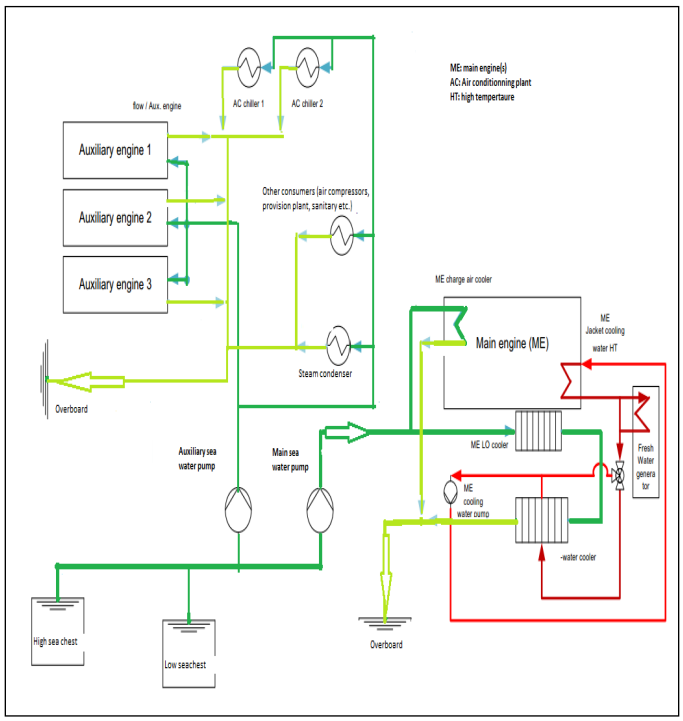
The inconvenient of this system are:
-The sea water temperature must be below 50°C to avoid scale formation.
-contamination of the water supply with consequent deposition inside the piping and cooled equipment.
- the system is subject to high rate of corrosion and erosion due to the nature of sea water.

2-Indirect cooling system with freshwate (Fig.2): machinery components are cooled by treated fresh water in a closed circuit. This fresh water is pumped through sea water coolers where its cooled by sea water.

This system reduces the length of sea water piping inside engine room and thus eliminates the corrosion problems linked to sea water. Furthermore, sea water pipes may be made from brass with limited costs.

Also, fresh water as cooling media permits the use of plate cooler instead of tubular cooler for their efficiency and easy maintenance. Figures 1&2 illustrate the simplified drawings of the two conventional cooling systems respectively direct and direct. Only main components are shown.

3-Keel / box cooler based cooling system: In this system, the cooling fresh water is cooled in coolers fitted outside the engine room. Sea water cooling pump and pipes are eliminated. This system is further explained in next chapters.



3. Description of keel cooler and box cooler.

The terms box cooler and keel cooler refer to the heat exchanges mounted on the underside of ship outer shell or in sea chests under the water line, saving space in the engine room.

Box cooler is a vessel cooling system, in the form of U-tube-bundle (Fig.3 and Fig.4) that is fitted in a sea-chest (Fig.5) on the side of the vessel. The sea chest is equipped with inlet and outlet-grids for circulating cooling sea water. The heat exchange takes place in the sea chest by natural convection of the water when the vessel is stationary or by a circulation due to the speed of the vessel.

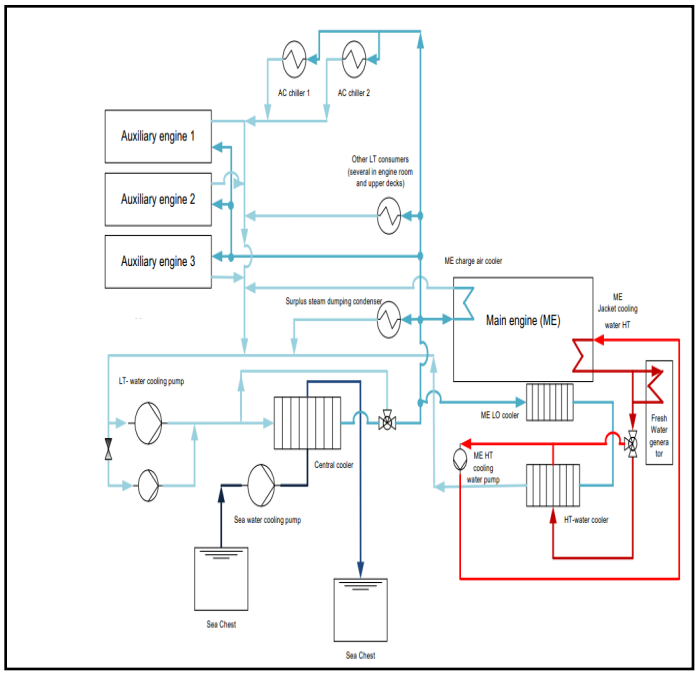


Figure 2. Indirect cooling system with fresh water.



Figure. 3. Box cooler: Main engine (left), Auxiliary engine (right). Photos taken during the vessel "ASD" stay in dry dock (2017).

Keel cooler (Fig.6): The closed circuit of fresh water is pumped in the spiral tubes installed outside the ship's hull below the water line, its contact with the sea water ensures the heat transfer and achieving the cooling effect. Compared to

conventional cooling systems, a keel / box cooling system provides several advantages.

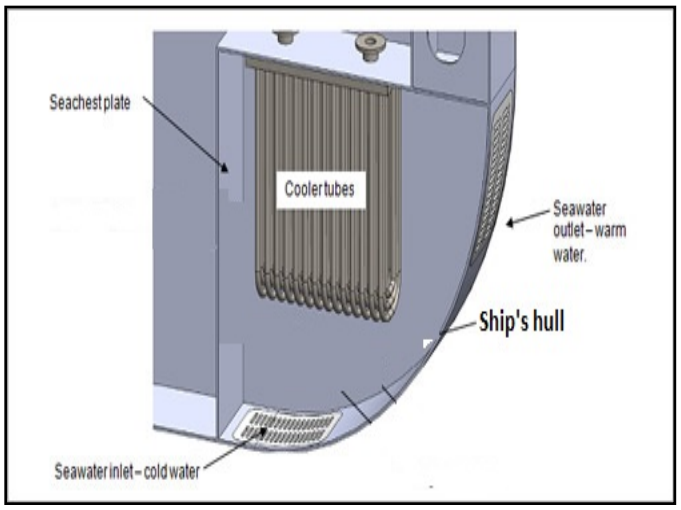


Figure 4. Box cooler and sea chest. (Google image).

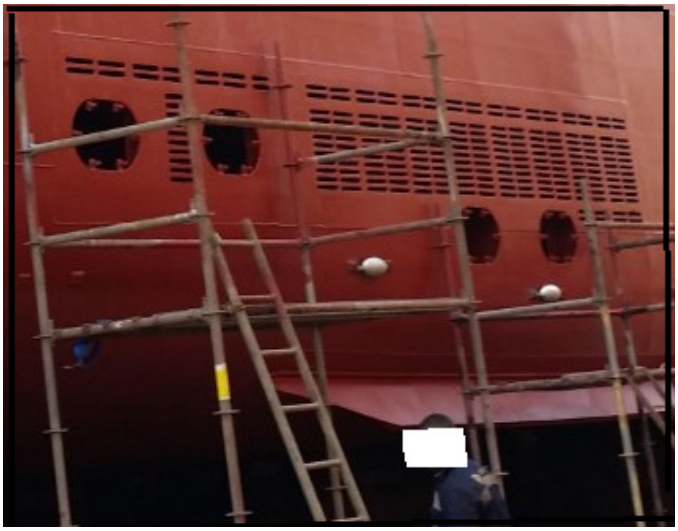


Figure 5. Sea chest where box coolers are fitted ("ASD" at dry dock 2017).

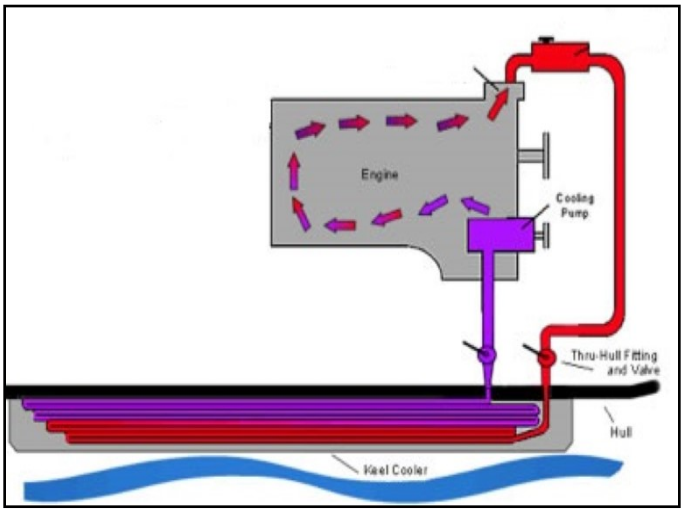


Figure 6. Basic keel cooling system. (Google image).

The sea water pumps, filters, sea water piping are eliminated, operational cost for cleaning, repairing and renewing the system components are reduced. It can also operate in all sea water conditions: icy, silt or polluted. However, the use of box and keel coolers raises some challenges, mainly the material corrosion and the biological fouling.

Nowadays the corrosion caused by box coolers may not a big problem due to technological advances in manufacturing resistant materials. The coolers are fabricated from noble materials such cooper alloys deemed corrosion resistant. The galvanic effect caused by material difference between the ship's hull made from steel and coolers is overcame by coating the sea chest surface and in some case by ICCP.

The biological fouling by barnacles, algae and other shellfish depends on the area where the ship is trading and operational conditions. Nowadays several solutions exist to prevent this problem: antifouling coating, electrolytic, chemical injections etc.

The use of box coolers is also limited by its capacity, according to makers, existing box coolers in the market may be used for cooling engines with output up to 30,000KW. Most of the commercial ship are fitted with engine of less power than this limit.

4. Case study

For this paper, we take for study the machinery arrangements of the vessel "Aline Sitote Diatta (ASD)", this vessel is a "Car Ferry" type used for the transport of passengers and vehicles between ports in Senegal. The particulars of this vessel are summarized in table I.The vessel is car ferry type fitted with two four strokes main engine (ME) with rated power of 1800 Kw each.

The electric power is produced by three sets of diesel generators 433 KW each. One generator can supply all necessary power during normal seagoing. The ship is equipped with two air conditioning units for all accommodations: passengers and crew cabin etc.As fuel, the ship main engines and auxiliary engines are burning marine diesel oil.

Table 1: The ship's particulars.

Length	76 meters
Speed	14.5 Knots
Gross tonnage	3481
Main engine output	Two diesel engine 1800 KW each
Propeller type	Controllable pitch propeller
Capacity	504 passengers, 28 vehicles

Figure 7 is a simplified representation of the machinery cooling installation on board the vessel "ASD". Set of box coolers are fitted in sea the chests on both sides of the ship. Each engine water coolant (fresh water) is circulated by engine cooling pump through one box cooler, where the coolant is further cooled by sea water before returning to the engine. Heat transfer is achieved by natural convection and circulation due to ship movement.

Compared to figures 1and 2, we can notice that main engine sea water pump and auxiliary engine pumps are eliminated including associated piping, valves and strainers.

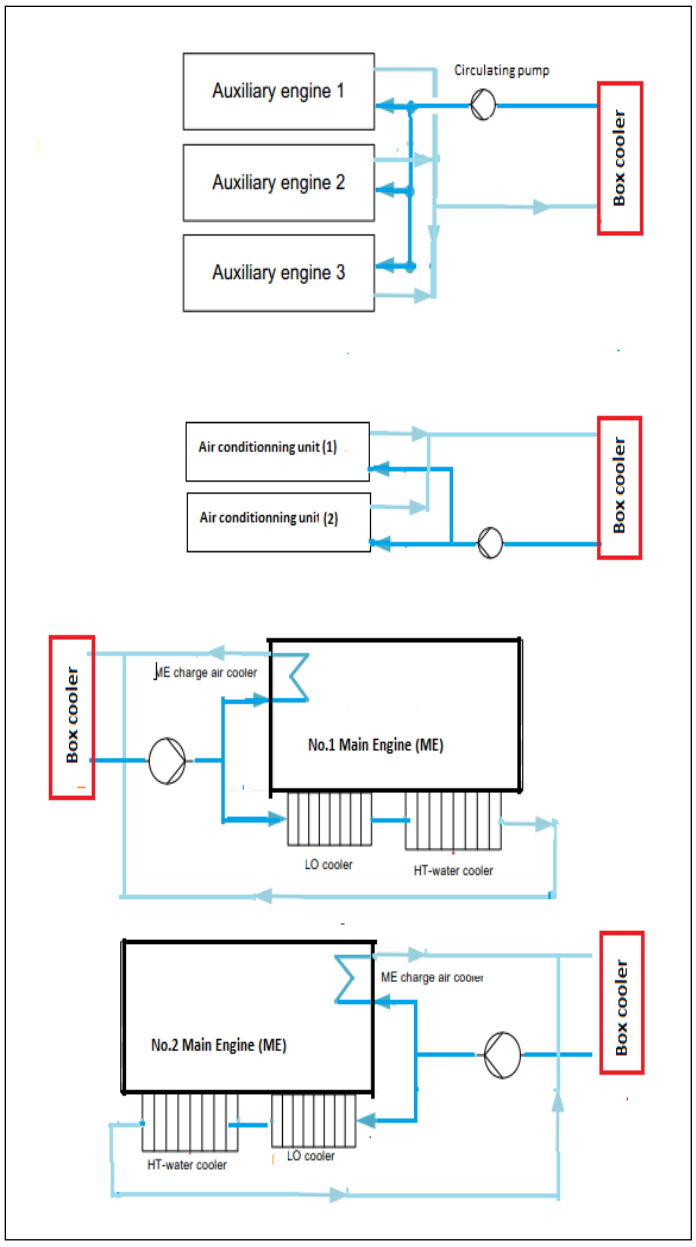


Figure. 7. Simplified represenation of the cooling diagram of the vessel "ASD".

4.1. Cooling system parameters calculation:

Determination of the sea water flow:

The quantity of sea water (V_0) expressed in *cubic meter per hour* (**m³**/*h*) necessary for cooling a single diesel engine is defined by formula (1).

$$V_0 = \frac{Q}{C_{sw} \cdot \rho . \Delta T} \quad (1)$$

Where
Q: Heat rejection of Diesel Engine in KJ/h
C_{sw}: Specific heat of sea water = 3.925 KJ/Kg˚C.
ρ: Sea water density = 1027 Kg/ m³
ΔT : Difference between the inlet and the outlet temperature of engine sea water cooling.

As per the engine specification published by the Maker, the total heat rejection (Q)the case study engine is [9]: Jacket cooling water, Hight Temperature (HT) circuit (375 KW); Charge air cooler, Low temperature (LT)-circuit (607 KW); Lubricating oil, LT-circuit (270 KW).

Note: Due to soft scale deposit inside the sea water piping and coolers which affect the water flow and cooler efficiency, the sea water cooler outlet temperature is generally maintained below to 50˚C. Taking in consideration that sea water temperature in certain navigation zones is up to 36˚C, we take $\Delta T \approx 14$˚C.

By substitution into formula (1) we can calculate the volumetric rate of sea water flow necessary for one engine cooling needs which is approximatively ≈ 80 m³/***h*.** The ship is fitted with two main engines. The ship is also equipped with three auxiliary engines for electric power generation but in worst case only two engines are sufficient to cover the ship power needs. Based on the auxiliary engine technical data, the sea water cooling flow necessary for cooling two auxiliary engines running in parallels is ≈ 30 m³/**h**.

As the ship is intended for passenger transport the air conditioning plant is important and its capacity is considerable hence the sea water flow required for the AC condenser as per its specification is 20 m³/**h.**

The total cooling sea water flow is the sum of:

(2x80+30+20) m³/**h**

Taking safety margin as 10%, The total sea water rate necessary for cooling the machinery at full load is approximately 230 m³/h.

Sea water pump power:

As per Bernoulli equation, the hydraulic power (p) (expressed in watt) transmitted to the sea water as it passes through the pump is calculated by formula (2)

$$p = V_s . H . \rho . g \quad (2)$$

Where
V_s is the sea water rate in m³/s, ρ is the density of sea water ($\rho = 1027$ kg/m³ at the sea surface), H is the total manometric height of the pump and g is acceleration due to gravity, average g=9.81 m.s⁻²).

The total manometric height H is defined by formula (3):

$$H = \left(\frac{p2-p1}{\rho . g}\right) + J_r + J_n \quad (3)$$

$p1$: pump suction pressure, in Pascal (Pa)
$p2$: pump discharge pressure, in Pascal (Pa)
Jr : geometrical height of the pumped medium lift, in meter (m)
Jn : head overall loss, in meter (m)

The difference in dynamic heights is negligible.

Jn includes all losses due to friction inside piping, flanges and valves. Its value depends on the fluid velocity, piping material, piping geometry etc.…

Sea water pump discharge pressure is globally around 2.5 bars (total H.≈25m), based on data published by marine diesel engine makers. By solving the equation (2) we define the hydraulic power absorbed by the pumped sea water flow of 230 m^3/h.

$$p.\approx 16 \text{ Kw}.$$

The sea water pump is driven by an electric motor, the electric motor power (P) is obtained from formula (4).

$$P = \frac{p}{\eta_e.\eta_m} \quad (4)$$

- η_m: pump efficiency, sea water pump is a centrifugal type. The average efficiency for such pump is 60%. [10]. η_e: is the efficiency of the pump electric motor approximatively 90%.

From formula (4) we defined the electric power necessary for driving the cooling sea water pump P= 16/06x0.9 ≈ 30Kw, equivalent to 1.7% of the main engine output. The specific fuel oil consumption for the modern four stroke engine at optimal service conditions is [7,11]:

$$SFOC = 200 \text{ -}210 \text{ g/Kwh}.$$

Therefore, eliminating sea water pump means saving 150 kg of fuel oil. This is equivalent to 450kg CO2 emission prevention per day. Considering that main engines are used for an average of 260 days per year, the total fuel oil save per year is 39 tons and 117 tons of CO2 emission prevention. The average price of marine fuel oil is 500 $/ton [12]. The application of box cooler for the cooling system of the vessel under study permits to:

-enhance the ship energy efficiency, by reducing the fuel oil consumption and CO2 emissions.
-reduce the ship operational cost.

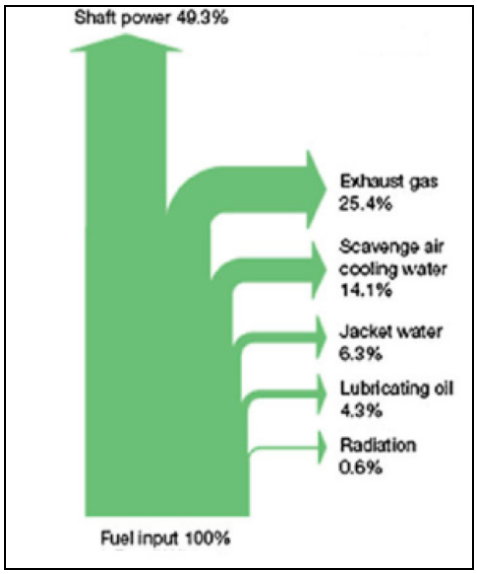


Figure. 8. Energy balance for typical marine diesel engine

4.2. *Typical diesel engine energy flux and cooling system energy*

The transformation of chemical energy to mechanical work by burning fuel oil is subject to various heat losses. The diagram in Figure 8 also called Sankey diagram, illustrates the basic energy balance of a modern Diesel. The heat losses by cooling engine lubricating oil, engine cooling water and the exhaust gases are mainly produced by the irreversibility of the engine cycle and only part of the energy is transformed into useful mechanical work.

The diagram in Figure 6, also called Sankey diagram, illustrates the overall energy balance of a modern Diesel engine, the values shown on the diagram are an average, slight deviations are possible depending on engine category, type, and power. Globally, the heat losses in marine diesel engines are estimated to 26-30%, this amount needs to be absorbed by the cooling system mainly with sea water.

Taking in consideration that the most used type of fuel for ship bunkers is the heavy fuel oil having a lower calorific value (LCV) of 42,700 KJ/Kg at ISO conditions. The specific fuel oil consumption for the modern two stroke engine is better than for the four-stroke engine.

SFOC = 160 g/Kwh.*The actual measured values for various engines is within range of 185- 205.*

The input quantity of chemical energy necessary per KWh output is given by formula (3).

$$Q_i = SFOC.LCV, \text{ (KJ/KWh)}. \quad (3)$$

The total fuel energy consumed by the engine per hour (input) is given by:

$$Q_t = Q_i.P \text{ , (KJ/h)}. \quad (4)$$

(P) *in KW*: the engine power transmitted to propulsion shaft (output). The heat losses Q_l to be evacuated by the cooling system is within 26% - 30% of Q_t (KJ/h). Figure 8 as example.

$Q_l \approx$ *26-30%* Q_t
Q_l (KJ/h) ≈ (26 – 30) % (0.160 x 42700 x P)
The heat to be dissipated in Kw is equal to Q_l/3600

From formulas (3) and (4) we can assume that the quantity of heat losses (in KW) to be dissipated by cooling system may reach up to 60% of diesel engine output. The cooling sea water rate and the electric power necessary for its pumping are defined by equations (1) and (2).

Table II is extracted from engine specification [11] it shows the cooling water capacity and power for modern diesel engines with different outputs. The column 5 is calculated based on SFOC of 200 g/Kwh, the optimal value for diesel generators. The table II covers only the main engine cooling system. The ship cooling system must be designed for also the auxiliary machinery components: diesel generators, steam condensers, cargo plant, air conditioning etc. The sea water flow and the pump capacity are higher than the values indicated, the fuel oil save and emissions preventions as well. TABLE II shows that the cooling sea water pump for single engine may absorb up to 0.4% of the engine output.

Depending on the size of this machineries, cooling system power needs may reach up to 1.7% as per the case studied.

Table 2: Propulsion power X Cooling power

Engine output (KW)	Heat dissipation (KW) (Maker data)	Cooling pump capacity (m3/h)	Cooling pump power (KW)	*Equivalent Fuel consumption per day for cooling (kg)
29000	17230	850	110	528
40670	24080	1177	153	735
58100	34480	1680	217	1042
81000	48000	2400	310	1488

World fleet outlook; Fuel oil consumption:

Table 3 is an extract from "DNVGL" report on 2030 world fleet outlook [14]. The total world marine fuel oil consumption is estimated to 325 Million tons per year equivalent to 1.028 Million tons of CO2 emissions.

Table 3: World fleet X Fuel oil consumption

Ship's type	Number of ship	Max. engine power	Average engine power	Average consumption per year (in tons)
Crude	2037	36,941	15000	10760199
Products	5272	20,080	2390	7242117
Chemical	3895	14,758	4185	14469811
LPG LNG	1725	39,902	8252	11293652
Bulk	7392	30,099	7830	49570494
General cargo	18473	16,550	2270	31986087
Container	4138	80,911	21880	63722049
Ro-Ro - Ferry	8859	52,799	4660	34889212
Cruise - Yacht	1550	75,627	6710	5945856
Offshore	5086	45,199	3821	4564645
service	17303	82,510	2162	14041725
Miscellaneous	24516	49,910	1060	15647195
	100246			**280965660.**

Total fuel oil consumed by shipping sector is evaluated to 325 million tons, 281 for propulsion and 44 consumed by auxiliary engines.

Reduction of 0.5% of fuel consumption by adopting box cooler means saving 1.6 million tons of fuel oil yearly. From the same table, we notice that less than 3% of ships have propulsion power of more than 30,000KW therefore 97% of world fleet may be fitted with central cooling based on box coolers.

5. Conclusion

The machinery cooling system of the vessel "Aline Sitoe Diatta" was studied. Two cases where considered, use of conventional cooling system and application of box cooler. Both cases are compared for energy efficiency and fuel oil consumption. The main conclusions are:

The ship cooling sea water pumps are the largest auxiliary pumps on board and are running most of the time therefore there are the largest electric power consumers, up to 10% of ship's needs in kilowatt – hour. For single engine, cooling water pumps absorbs as a minimum 0.4% of the engine output.

In the case studied cooling system of all main and auxiliary machineries requires up to 1.7% of the propulsion engine power. In the opinion of authors, the use of the keel or box coolers for ship machinery cooling system may contribute to reduce fuel oil consumption of the ships, to improve the energy efficiency accordingly and meet the legal requirements [13].

The application of box cooler is limited to small size ships but further to recent technology developments it's possible to adopt on ships up to 30,000 KW (only 3% of the world cargo vessel is equipped with engine of more than 30,000 KW, mainly container vessels).

With large power, the box coolers may be part of the solution which will contribute to reduce the size and the cost of the cooling system and energy consumption.

References

[1] C. T. Wilbur "Pounder's Marine Diesel Engines" Sixth 6th edition.

[2] Chun-Lien Su, Wei -Lin Chung, Kuen-Tyng Yu. « An enegy saving evaluation method for adjustable frequency drives on sea water cooling pumps on ships. Industrial and commercial Power system techniques. 2013/ IEEC/IAS 49th

[3] Gazi Cocak, Yalcin Durmusoglu "Energy efficiency analysis of a ship's central cooling system using variable speed pump". Journal of Marine Engeneering and Technology: Published on 2017-01-31.

[4] Mia Elg, Maunu Kuosa, Markku Lampinen, Risto Lahdelma, Panu Mäkipeska, Juuso Raita, Guangrong Zou, Kari Tammi "Advanced auxiliary cooling system for energy efficient ships". Technical Research Centre of Finland Ltd. -C._T._Wilbur

[5] Gerasimos Theotokatos, Konstantinos Sfakianakis and Dracos Vassalos. "Investigation of ship cooling system operation for improving energy efficiency" Department of Naval Architecture, Ocean and Marine Engineering, University of Strathclyde, G4 0LZ, Glasgow, UK.

[6] Application of titanium in shipboard sea water cooling systems. Wayne L. Adamson Vol.99, Issue 3.

[7] Ait Allal Abdelmoula, Khalifa Mansouri "Toward a reliable sea water central cooling system for a safe operation of autonomous ship"/ ENSET Mohammedia, University Hassan II Casablanca.

[8] Andrzej Młynarczak. Box coolers as an alternative to existing cooling systems Scientific Journal. Maritime University og Szczein . 2013, 36(108) z. 2 pp. 131–136.

[9] Wartzilla 20 product guide. Issue 1/2017

[10] Sulzer_Centrifugal_Pump_Handbook,_3rd_Ed.M.

[11] MAN B&W S90ME project guide ed. 05/2014

[12] https://bunkerindex.com/

[13] IMO. (2011). Resolution MEPC.203(62), Amendments to the annex of the protocol of 1997 to amend the international convention. http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/Marine-Environment-Protection-Committee-(MEPC)/Documents/MEPC.203(62).pdf

[14] World fleet MACC 2030 Ver.24 DNV. 2012-02-12 GL Library..

[15] DNV GL SE, (2015). Rules for Classification and Construction, I-1.2, July 2015, Hamburg, Germany DNV.GL website publication 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 83-98 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Theory Building of Quattro Bottom Line Approach for Sustainable Reverse Logistics from Government Perspective: the Indonesia Evidence

Hesti Maheswari*, Gatot Yudoko, Akbar Adhiutama

School of Business and Management. Institut Teknologi Bandung, Bandung 40132, Indonesia

ARTICLE INFO

Article history:
Received: 11 June, 2018
Accepted: 7 July, 2018
Online: 23 July, 2018

Keywords:
Sustainable reverse logistic
Quattro bottom line
Government perspectives

ABSTRACT

Reverse logistics activities are widely used application to preserve the environment in many developed countries from e-waste problem. Due to lack of public awareness in many emerging countries, implementation of reverse logistics program must be initiated by government. Besides that, reverse logistics practices often fail to balance among profit target achievement, the environmental conservation, and social harmony. Moreover, in the recent years, triple bottom line is not enough since many parties are only focused on their reputation due to they are afraid to be blame as the source of problems. Therefore, this research aims to build sustainable reverse logistics theory for emerging countries through quattro bottom line approach from government perspectives. Theory building process consists of systematic literature review, triangulation of data collection, major part of theory description, conceptualization and operationalization of theory. The result indicates a number of obligation that must be regulated by government for all parties that was divided into four categories, i.e. infrastructures and facilities provision, human resource commitment, regulation establishment, and managerial orientation. This theory can be adopted to measure performance of SRL Implementation.

1. Introduction

This paper is an extension of work originally presented in Industrial Engineering and Engineering Management Conference, IEEM 2017 [1] that wants to prove the need for government role in building sustainable reverse logistics (SRL) theory by using quattro bottom line approach. This study departs from the researcher's belief that the success of handling e-waste depends on an active role of government. At least there are three roles of government in supporting reverse logistics (RL) implementation. First, as a policy maker, government makes regulations and policies in preserving the environment [2], second, as a regulator, government arranges RL implementation, and third, as a facilitator, government provide facilities to conduct RL activities. Actually, electronic and electrical products are considered as luxurious and unaffordable things for middle-lower income country. However, due to the increasing of the Indonesian income per capita that almost reached 4 million per month, communities consumptive behavior, and the growth of industrial capability reflected in gross domestic products (GDP) of Indonesia more than 5%, the third highest in the Group of Twenty (G20) after China and India [3], lead the community to replace electronic products faster than it should be. Generally, the Indonesian people persist on using their mobile phone (MP) are not more than three years [4], which causes these products contribute the largest amount of e-waste compared with other electronic devices [5].

Along with the highest rate of electronic products replacement, increase also the amount of e-waste, while, only a few of them are recycled. Nokia mobile phone survey noted that only 9% of e-waste mobile phones are recycled in India and China, and only 1% in Nigeria and Indonesia [6]. The biggest role in recycling activities comes from informal businesses [7] that pay more attention only to target profit achievement and do not care with the environment. This is the reason for the great role of government to control recycling programs. Through government policies and regulation, not only balancing profit and reputation achievement, but also protecting the environment and social harmony. The rules and regulations should be designed properly to arrange how to take benefit from recycling process without disturbing other interest.

As mentioned before, the success of a a country in conducting recycling program depends on the role of government to encourage

*Hesti Maheswari, Jl. Rudal Raya No. 3A Joglo West Jakarta 11640 - Indonesia, +62 82123596058 email: hesti.maheswari@sbm-itb.ac.id

and regulate it [2]. Providing a high technology recycling needs enormous budget allocation. In developed countries, the amount of e-waste is regarded as a resource for other business. They convert e-waste into energy or other valuable products. Furthermore, the government early anticipate the number of product produced in forward chain. They think early to create a handling mechanism by making RL activity as an obligation for everyone, either businesses or societies. Unlike in developed countries, regulations in emerging countries are always left behind. The rules or acts will be created after government feels there is a crucial problem. Even though RL activity is conducted, they are often ignoring the environmental friendly way. The main characteristics of emerging countries RL model as compared to international models [7] is depicted in Table 1.

Table 1: Comparison of RL models in developed and emerging countries

	Developed Countries		Emerging Countries		
	Switzerland	Sweden	China	India	Brazil
Government	Active	Participative	Inactive	Inactive	Active
Industry	Total Mgt.	Transportation	Ships WEEE		Total Mgt.
Consumer	Pays ARF returns	Return			Pays embedded fee and returns
Informal S			Excluded	Pre-processing	No training

There is no opportunity for informal sector in developed countries to get profit through RL activities. The government strictly prohibits informal sector to recycle used electronic products because they are often ignoring the environmental sustainability. While, in many emerging countries, the informal sector plays a major role in recycling e-waste activities, even becoming home industry but illegal.

According to my preliminary research, in fact, there are many Indonesian people want to participate in recycling MP program by sending it to collecting point [4]. Unfortunately, most of them said that they do not know where to bring it for recycling since they are hard to find the drop box facility [4]. Some of them are reluctant to send their used MP, and prefer to keep it at home due to unreasonable of reimbursement. They are also hard to find someone else who really want to use it. Others explained that it is very difficult to sell the obsolesce or out of date electronic product to secondary market [4]. These phenomena show that recycling mechanism in many emerging countries that actually can be used for handling e-waste problem, are not built yet.

From literature reviews, we found that since 1990s until now, people all over the world have felt the need for RL, although their aims are not purely to protect the environment. They are focus on profit or embarrassment if their activities are littering surrounding [8, 9]. Therefore, in that era, many researchers studied about how to improve firms' performance in reverse chain. Evolution of RL theories refer to original constructs built by some previous studies such as reverse logistics network that explored the types of RL activities [10] and motivation [11], organizational commitment [11-14], and RL performance [15]. The author is hard to find a research that explores the role of government in encouraging communities to participate actively in RL program, how the government regulates RL mechanism, how the government assigns businesses to conduct RL activities, and how the government oblige societies to participate. Therefore, today it is relevant to sift the research idea to focus on government role in preserving the environment through RL program that consider not only profit, planet, and people elements, but also consider reputation factor. Due to low of public awareness in emerging countries even they pollute the environment although they conduct RL activities, therefore, the researcher believes that RL concept is not enough and this study proposes to use sustainable reverse logistics (SRL) terminology. Moreover, the researcher believes that SRL concept must be explored and created from government policies, regulations, perspectives and expectations. So, by investigating the linkage and the gaps among theories, finding the governments expectation by interviewing them, and generating a suitable pattern of the program, this study wants to build a SRL theory based on quattro bottom line (QBL) approach, include reputation terminology as a new idea and very relevant with current situations but has not emerged in the research publication.

The rest of this paper is organized as follows. The second section explores literatures related to sustainable reverse logistics, government rules, policies, and performance. The third section discusses the research method, data collection process and highlight the variables for building a theory. The fourth section explains the main finding. Finally, the fifth section summarizes the research and suggests further research.

2. Literature Review on Sustainable Reverse logistics

Exploration of sustainable reverse logistics concepts are directed to prove that the main purpose of conducting SRL programs are protecting the environment. Furthermore, the success of this program particularly in emerging countries, depends on the role of government to encourage other stakeholders. Therefore, the actions that support RL implementation, governments' performance, and sustainability concept will be the main themes in this paper. Related to sustainability concepts, this paper uses quattro bottom line approach because the researcher believes that it is not enough without spelling out reputation concepts for research in the context of emerging countries.

2.1. Sustainable Reverse Logistics Actions

Reverse logistics is the process of planning, implementing, and controlling the efficient, cost effective flow of used electronic products from the point of consumption to the point of origin for the purpose of recapturing or creating value to prevent environmental pollution [11]. From this definition, it is illustrated that RL activities should be able to protect the environment and one of the ways is adding span of life electronic products especially for MP that has the shortest product life cycle.

There are 8 types of reverse logistics activities [16], namely remanufactured, reconditioned, refurbished, re-process, repair, recycle, reuse, and disassembly. However, only 5 types are commonly done in emerging countries, i.e.: refurbished, repair, recycle, reuse, and disassembly [17] and usually many informal businesses are doing these activities. On the contrary, only a few manufacturers take a part in RL program and only a little role of government in regulating mechanism of RL implementation. Therefore, this research seeks to find actions that support SRL theory building, either from literatures or practices.

This research divides type of reverse logistics actions based on who is recommended by government to perform the activity, whether businesses (manufacturer or intermediary business), consumers, or governments itself. Knowledge, willingness, and awareness of all stakeholders to the environmental sustainability are believed to be a strong foundation to commit in supporting SRL program.

2.2. Sustainability

In manufacturing industries, after the achievement of quality, differentiation, and cost leadership as the leverage of their competitive advantage, management has to concern about strategic issue such as the environment. The most integrated performance of supply chain either in forward or reverse network is the business development in line with sustainability. Sustainability is a set of skill and concepts that allow all parties to obtain competitive returns without sacrificing people and environment interest [8]. Aptly joining the definition of reverse logistics and sustainability brings about sustainable reverse logistics (SRL). The industry must control not only financial achievement, but also consider its impact on economic as a whole, environmental protection, and social responsibility which conform to the triple bottom line (3BL): economic value, environmental and social responsibility (Figure 1) [18]. Others stated that the most important reasons in implementing RL activities are economic (decreasing the use of raw materials, reduction of disposal cost, creation of added value for and-of-use products, and demonstrating of environmentally responsible behavior), legal (implementing extended producer responsibility as follow the government rule), and social (preventing environmental pollution from carbon emission and e-waste generation) [19, 20]. The worries on corporate image, government responsibility, and communities' awareness are sometimes become the reason of all stakeholders to conduct RL programs.

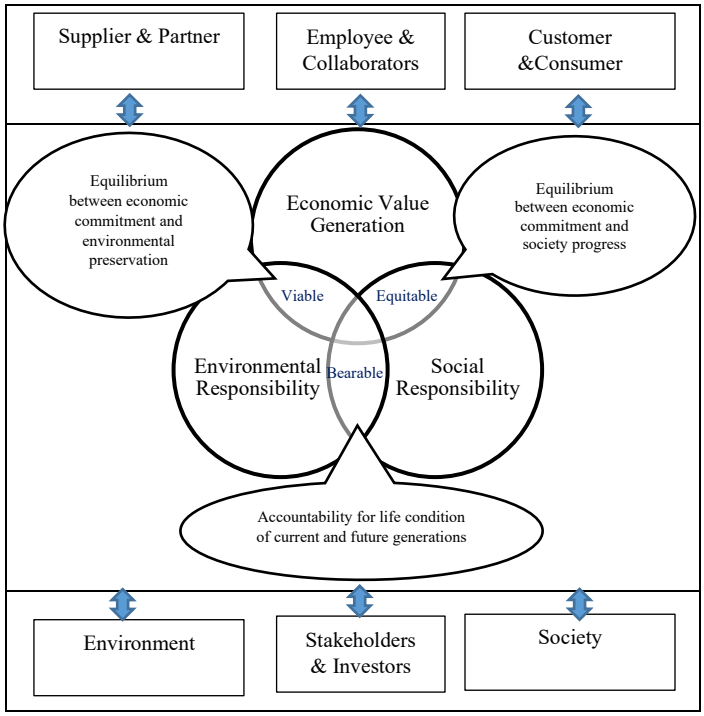


Figure 1: The triple bottom line [Pedroso in [16]

Figure 1 explains about interactions among social, environmental and economic spheres result in deeper and more complete aspects of sustainability. The relation between economic and social prioritizes job creation and income to improve living standards, economic welfare and growth that favor social aspects, besides increase level of education and professional development. Intersection between the economic and environmental spheres stresses to create benefit from various activities but still pay attention to environmental sustainability such as designing eco-friendly products and developing methods for electronic product volume reduction [21]. Further, interaction between social and environmental spheres indicates that communities have to use electronic products properly reflected in accountability for life condition of current and future generations. Finally, intersection of the three spheres represents sustainability in a broader sense. These issues will be the central topic in this study because most parties are often fail to maintain the balance among the profit target achievement, planet conservation, and social harmony.

2.3. Reputation

Reputation in an organizational context, is generally related to two dimensions, first, how stakeholders perceive the quality of specific attributes, and second, the degree to which an organization or company is recognized collectively [22]. From a business perspective, reputation has been defined as the combination of a company's public prominence, its public esteem, and the qualities or attributes associated [22, 23]. In other word, a company or organization is recognized on a large-scale and salient in the minds of stakeholders. From government perspective, reputation concepts are defined as responsibility of a country from various problems especially that giving negative impacts and disturbing to other countries. While, for the communities, reputation is related to their willingness to actively involved in preserving the environment.

There are six dimensions of reputation [24], namely:

- Emotional appeal (likeability, respectfulness, and trustworthiness);
- Physical appeal (attractiveness of an infrastructure);
- Financial appeal (favorable environment for stakeholders);
- Leadership appeal (charismatic leadership and a clear vision);
- Cultural appeal (socio-cultural diversity, history, entertainment);
- Social appeal (the perceived responsibility as a member of the global community and the manifest support for good causes)

2.4. Government Policy, Regulations, and Performance

Government has a range of policy instruments that can be divided into six broad categories to reduce the environmental impact of freight logistics [25], namely: taxation; financial incentives; tariff regulation; liberalization and privatization of freight markets; infrastructure and land using planning; as well as advice and exhortation. Government as a regulator can run e-waste management system through realizing extended producer responsibility (EPR) implementation; corporate social responsibility (CSR) mechanism, energy and row material control, take back program application, environmental-friendly product utilization, e-waste budget allocation, deposit fund requirement, international end-of-life standard, 4Rs program application [21]

and skill accreditation [26, 27]. Government as a facilitator should provide drop box that easy to find by communities, storage point, and landfill areas. Indonesia Waste Forum (IWF) requires five aspects which must be considered to solve waste problem in emerging countries they are law and national policy, national institution for integrated implementation, financial investment, community involvement, and technology creation [28].

3. Research Method

3.1. Method

The perfect research design is built to ensure all research questions will be answered [29]. This research uses quantitative and qualitative approaches and the design is divided into five steps that depicted in Figure 2.

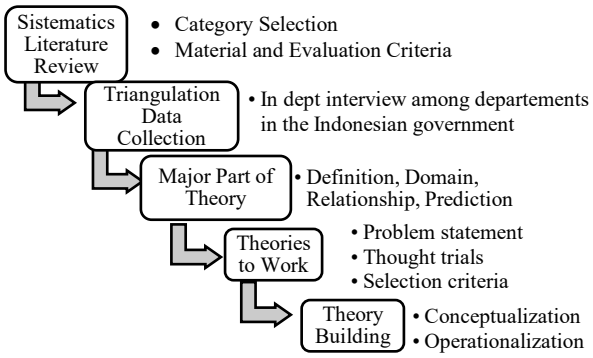


Figure 2: The research flow

First, in quantitative approach, this research chooses systematic literature review [30], that comprised of two steps i.e.: category selection and material evaluation. In the category selection step, firstly this study defines and refines a topic, then designs a search and locate research report. This study uses two main topics namely: reverse logistics and sustainability. This study explored many researches without limiting the year of their publication. After that, they are reviewed and classified by differing their characteristics and distinguished and then recorded in a prepared spreadsheet to be analyzed holistically. The different characteristics in reverse logistic concept are its activities and network while the sustainability concept refer to the fulfillment of Quattro bottom line approach. Secondly this study determines how to evaluate the material by using validity test, crosschecked for ensuring the enrichment of study such as Scopus indexed, journal accredited, and the research quality that answering its research question.

Because this research wants to build SRL theory from government perspective and expectation, therefore in the end of this section will be served the measurement of government performance that conduct RL activities. For evaluating performance of government, this research chooses three types of evaluations [31], namely: criterion-focused evaluation (by what criteria shall the SRL program be evaluated?), impact focus (what are the direct and indirect SRL program impacts, not only on participants, but also on larger systems and the community), and implementation focus (to what extent was the program implemented as designed? What issue surfaced during implementation that need attention in the future?)

Second, to complete the constructs, this research applies a qualitative method by using triangulation data collection (i.e. in-depth interview between departments in the Indonesian government) for six months from January 2017 to April 2018. *Third*, this research adopts that theory comprises of four major parts, i.e.: definitions, domain, relationships, and predictions [32]. Definitions translate the concepts, constructs, or variables used in the theory by using "what?" and "who?" questions. Domains refers to where the theory applies with answer the questions of "where?" and "when?". Relationships shows the nature of causality among those concepts, constructs, or variables selected by answering "why?" and "how?" questions. Predictions reflects whether the theory will give impact in the field by applying the questions of "could, should, and would?" *Fourth*, Weick's theorizing as disciplined imagination provide three tools for theorist to work with in crafting theories, i.e.: problem statement, thought trials, and selection criteria [29]. and *Fifth,* the general method of theory building in applied discipline comprises of five nonlinear phases: conceptualization, operationalization, confirmation, application and refinement [29]. This research is only use the first two phases, i.e.: conceptualize and operationalize. The purpose of conceptualization phase is to develop a sound conceptual framework that provides an initial understanding and explanation of the nature of problem and phenomena. While, the purpose of operationalization phase is to bridge the concept and practice. Although, the operationalization of a theory needs to be tested or confirmed, in this paper, the confirmation phase is not conducted because of the interview results refer to government policies and regulations. In turn, the theory is likely to result in better outcomes and appropriate to be implemented.

3.2. Samples

Since lack of community awareness in emerging countries to reduce e-waste and preserve the environment, so the theory formation must be generated from government perspectives.

Table 2: Government and selected respondents

National Agencies	Selected Respondents
• Ministry of Environment and Forestry (MEF)	• The head of sub directorate of e-waste and recycle • Directorate General of Waste, E-Waste, Hazardous and Toxic Material
• Ministry of Industry (MI)	• Directorate Industry of Electronic and Telematic • Directorate Small Medium Industry of Metal, Machine, Electronic and Conveyance
• Ministry of Trade (MT)	• The Head Agency of Assessment and Development of Trade Policies
• Ministry of National Development Planning (MNDP)	• Directorate of Environment
• Ministry of Cooperation, Micro and Small Medium Enterprise (MCSMEs)	• Secretary for Deputy Minister of Business Restructuring
• Environment and Hygiene Office (EHO)	• The Head of Environment and Hygiene Office • The Head of Hygiene Management • The Head Control of Environmental Impact
• Cooperation, Small Medium Enterprise, and Commerce Office (CSMEs C)	• The Head of Cooperation, Small Medium Enterprise, and Commerce Office

By interviewing the government officials, this research wants to catch the various constructs considered for implementing SRL activities. This research chooses five ministries and two government agencies that are predicted influencing the successful of SRL implementation. Not all of divisions in those ministries made as respondent. Table 2 is detailing who the respondents are.

4. Result and Discussion

This section consists of five parts following the research flows. The first part is exploring the existing theory include the evolution of RL and further about SRL theory. The second part is recording the government policies and regulation plus roles and responsibilities from in dept interview. Then, the third part is elaborating the definition, domain, relationships, and prediction of SRL theory, the fourth part is selecting the SRL criteria through problem statement and thought trials, and the last part is pattern matching of the concept with the practices to create the sound of sustainable reverse logistics theory.

4.1. Evolution of Reverse Logistics and Sustainable Reverse Logistics Theory

Firstly, some businesses conducted reverse logistics activities for getting profit from used electronic product [4, 11, 12, 15, 18, 33-35]. Actually, the idea of returning electronic product emerged around 1940s to 1970s. They did not recycle or conduct other form of RL activities. Usually they resell used electronic products to other. If those products cannot be sold, they kept it at home, sometimes become a children's toy, or throw them in the bin mixed with household waste. However, in the recent years, they began aware with environmental condition which resulted by more and more used electronic product that make the narrowness of dumping grounds. Actually, RL activities are closely related to sustainability concepts with the three perspectives (triple bottom line, 3BL), namely: profit, planet, and people [8]; economic value, environmental and social responsibility based on Pedroso and Swicker [18]. Types of reverse logistics activities such as remanufactured, reconditioned, refurbished, re-process, repair, recycle, reuse, disposal, and disassembly [10, 15, 18] are manifestation of sustainability movement. However, for many emerging countries, the use of 3BL concept is not enough. RL managed efficiently and effectively has the potential to gain economic value and enhance the positive image of the firm in the consumers' perception and the distribution chain [33]. RL managed effectively help to improve after-sales service as fast to respond to complaints and solve problems which is able to provide certainty that the returns enhance the positive image of firms [20]. The firms deliberately doing RL activity to get attention from societies because mitigating environmental, health, and safety impacts can enhance a firm' reputation [8]. For the government, they worry if their e-waste pollutes neighboring countries and they get negative image from them. Today, societies also worry about getting social sanction, if their behavior is not responsible or lack awareness to the environment. Social perspective to assess reverse logistics performance is also considered as part of reputation concept with some indicators such as market reputation of the enterprise if they produce unqualified product and do not have a technology for handling their own e-waste [15]. This matter makes the researcher believe that not only businesses but also government and communities expecting a good reputation in realizing it [4].

Discussion of sustainability are driven by the basic concept that reverse logistics' performance should be measured not only by profits and reputation, but also by the impact of activities on environmental and social systems. Therefore, there is a tight correlation between reverse logistics and sustainability. Unfortunately, in many emerging countries, in fact, RL activities are often not managed and controlled properly by government. They are still concern to produce a new qualified electronic product in forward chain, provide technology to develop features, promote, and sale it in the market as much as possible. They forget to early anticipate e-waste problem. Meanwhile, the informal sector from community more quickly respond by looking at it as a business opportunity. After that, a new problem emerges in almost all emerging countries as they move without government permission and ignore health and safety work. They only think about how to fulfil their needs or how to earn a lot of money. These phenomena make the author convince to use sustainable reverse logistics terminology if we want to study about all activity in reverse chain.

4.2. Government Policies and Regulations plus Roles and Responsibilities

Many emerging countries have no rules, act, or legislation completely related to SRL implementation. Some rules bellow will be used as a foundation of building SRL theory. For example, Indonesia just made a rule about the obligation of businesses to manage their used product, if it is hard to decompose by natural process in Act No. 18 of 2008, article 15.

The last government regulation related to face e waste problem is No 101 of 2014, article 1, verse 11 – 31 that concern with managing e-waste. Indonesian government also prohibit businesses to import used electronic product, e-waste, hazardous, and toxic material as shown on Decision of The Ministry of Trade No. 229/MPP/Kep/7/97 and Basel Convention of 1989. Polluter pay principle is also set in Act No. 32 of 2009 article 2 verse (i) and Government Regulation No.18 of 1999 article 9 verse (1, 4, 5). For reducing e-waste, government assesses and control the quality of product through national standard of Indonesia (SNI) that enforced mandatory for electronic product in regulation of The Ministry of Industry No. 36/M-IND/PER/9/2017. Indonesian government form adjusted assessment institution in order to SNI enforcement.

Ministry of Industry has machines restructuration and/or other equipment program for small and medium industries in rule of Ministry of Industry No. 20/M-IND/PER/3/3/2016to help them improve the quality of products. So, the procurement of machine should not be a problem anymore for converting e-waste to other valuable product. Government also encourage all businesses to reach green industry certification as arranged in rule of The Ministry of Industry No. 18/M-IND/PER/3/2016. For increasing the number new of entrepreneurs, Indonesian government makes new business invention program and train them until it becomes an established business as shown on Act no 20 of 2008. Government make serious effort to link and match small and medium enterprises with the big industry to learn the technology and improve their skill that arranged in PP No. 17 of 2013. There is no rule, act, and other form of regulation that organize how to conduct SRL activities, who should be responsible, what the role

of each actor is, and what kind of punishment if they are breaking the rule. As explained in previous section, because of community lack awareness the concepts of SRL must derive from government initiative. Therefore, this author formulates the theory of SRL from the government perspective.

This research explores the roles and responsibilities of national agencies related to SRL implementation by interviewing them. These roles and responsibilities will be a fundamental construct in building SRL theory (Table 3)

Table 3: Government roles and responsibilities

National Agencies	Roles and Responsibilities
• Ministry of Environment and Forestry (MEF)	• Define responsibilities for stakeholder • Define SRL categories for business • Establish and manage e-waste collection channel and systems • Establish e-waste treatment standard • Manage licensing system for business • Monitor and evaluate business' environmental performance
• Ministry of Industry (MI)	• Encourage eco-design product • Finding RL technology • Create technology e-waste conversion
• Ministry of Trade (MT)	• Prohibit import of used electronic product. • Define product quality
• Ministry of National Development Planning (MNDP)	• Develop financing scheme SRL • Define and manage subsidies
• Ministry of Coop. Micro and SME (MCSMEs)	• Training the business • Create link and match with big industry

4.3. The Major Part of Sustainable Reverse Logistics Theory for Emerging Countries

The SRL theory especially for mobile phone industry has not been extensively studied and conceptual models explaining the possible recovery network have not been abundantly created [8]. So, the implementation of RL activities in almost all emerging countries is more damaging the environment because it still does not consider sustainability concepts since there is no theory that fortifies these activities. For gaining the sound theory, four major parts of theory (definition, domain, relationship, and prediction) [32] will be described in table 4-7. This research explains what SRL is, who should play a role, when and where it must be implemented, why it is important, how it is applied, whether it can be implemented, how it should be implemented, and will all parties participate.

Table 4: Definition component

Questions	Description
What?	Sustainable reverse logistics is designed as an attempt to guarantee activities in reverse chain in line with the primary goals such as profit, social harmony, and environmental preservation. This theory is really needed especially in emerging countries to fortify community activity in utilizing used electronic products to get profit.
Who?	Governments are the backbone in enforcing sustainable reverse logistics implementation who formulate the mechanism includes educating and socializing to community and also organizing it. Due to implementation of sustainable reverse logistics theory requires collaboration of all parties, therefore not only government, but also manufacturer or producer, intermediary business, and society should be involved

Table 5: Domain component

Questions	Description
When?	Sustainable reverse logistics should be done when production of electronic products in forward chain takes place and followed by consumptive behavior, to anticipate e-waste generation
Where?	Sustainable reverse logistics is designed for community in emerging countries especially in a country where the population has low of the environmental preservation awareness

Table 6: Relationships component

Questions	Description
Why?	The Indonesian government, through Ministry of Environment and Forestry, Ministry of Industry, Ministry of Trade, Ministry of National Development Planning, and Ministry of Cooperation, Small Medium Enterprise should concern with the environmental quality degradation due to a lot of e-waste. Because of lack community awareness, the sustainable reverse logistics theory must be developed from government side. Government policy and intervention will be the right breakthrough for encouraging all parties aware with environmental sustainability. Finally, this theory is expected become a guidance for all parties who are looking for profit on reverse chain.
How?	The Indonesian government can apply sustainable reverse logistics theory by arranging its mechanism, making the rules and regulations, determining the role of each stakeholder, developing infrastructures or facilities, preparing the community knowledge, and controlling the implementation continuously. The expected outputs from this theory building refer to sustainability achievement through quattro bottom line approach, encompass business profit either for producers or intermediary businesses, local and national economic growth, social welfare, and environmental preservation as well as all parties' reputation since their effort and participation.

Table 7: Prediction component

Questions	Description
Could?	The Indonesian Government believes and expects that sustainable reverse logistics theory could be implemented in emerging countries for handling e-waste problems.
Should?	To implement sustainable reverse logistics theory, the activities should refer to quattro bottom line fulfilment. The balancing among profit target achievement, social value creation, environmental preservation, and reputation acquisition should become a fundamental orientation.
Would?	All parties would adhere to government regulation since they receive equal treatment between the fulfilment of rights and the implementation of obligation

4.4. The Categories Selected for Sustainable Reverse Logistics Theory for Emerging Countries

The SRL categories combining literatures, field practices, and also government roles, responsibilities, regulation and expectation are summarized in table 8. This research found 53 categories which will be used as the constructs for SRL theory, while the definitions of each categories are provided in the appendix. There are 17 constructs are not found in literature, but regarded very important in building SRL theory for emerging countries. They are standard equipment utilization; training for SMEs in reverse logistics activities; inclusion of product content and duration; fairness in using electronic product; illegal import restriction, Electronic product national standard arrangement; vocational education for human resource preparation; business localization; tax arrangement; restructuration machine program; budget provision; rules alignment; social punishment; control

mechanism of RL implementation; SMEs-industry link and match; Modernization of SMEs; and social and public facility utilization. We give star mark for the seventeenth constructs to differentiate the both constructs source clearly.

Government recommends 18 actions of SRL for businesses. Three of them are found from interview result, namely encouraging utilization of standard equipment to avoid hazardous and toxic materials from e-waste; giving training to small medium enterprise in e-waste processing and use of e-waste conversion technology, and writing down the product life and product content on the packaging. Although other 15 action are found from literature, however only 2 actions have been conducted in Indonesia, i.e. giving after sales service and providing drop box facilities but with limited number.

Table 8: Sustainable reverse logistics concepts for emerging countries: summary of in depth interview results and literature review

Sustainable Reverse Logistics Categories	Actions Recommended by Governments							Literatures
	MEF	MI	MT	MNDP	MCSMEs	EHO	CSMEC	
Businesses:								
Eco-friendly product design	√	√	-	-	-	-	-	[20]
After sales service	√	√	√	-	-	-	-	[38]
Packaging reduction	√	-	-	-	-	-	-	[39]
Deposit fund	√	-	-	-	-	-	-	[26]
Take back obligation	√	-	-	-	-	√	√	[9]
Drop box facility	√	-	-	√	-	√	√	[5]
Recycling technology creation	√	√	√	√	-	√	√	[16, 26, 38]
Standard equipment utilization*	√	√	-	-	√	√	√	-
Skill accreditation (employee competency)	√	√	-	√	√	√	√	[14]
Quality of product (longevity of extension)	√	√	√	-	√	√	-	[26]
Organizational commitment	√	√	-	-	-	√	-	[11, 35]
Budget allocation	√	√	-	√	√	√	√	[13]
Inclusion of post-use product handling	√	-	√	-	-	-	-	[34]
Technology conversion	√	√	-	√	√	√	√	[14, 26]
SMEs coaching*	-	√	√	√	√	-	√	-
Enterprise alliances	-	√	-	-	√	-	-	[14, 36, 38]
Inclusion of product content and duration*	√	-	-	√	-	-	-	-
Corporate citizenship	-	-	-	-	-	-	-	[41]
Public:								
Public participation awareness	√	-	-	√	√	√	√	[14, 9]
Eco shopping decision	√	-	-	-	-	√	-	[20]
Reduce replacement frequency	√	-	-	√	-	√	-	[16, 26]
Separate own e-waste	√	-	-	√	-	√	-	[40]
Reasonable reimbursement (incentives)	√	-	-	-	-	√	-	[8, 40]
Willingness to pay recycling cost		-	-	√	-	-	-	[40]
Fairness in using electronic product*	√	-	-	-	-	√	-	-
Government:								
Illegal import restriction*	√	-	√	-	-	-	-	-
Electronic product national standard arrangement*	-	√	√	-	√	-	-	-
Collecting point extension	√	-	-	√	-	√	√	[26]
Invention of handling mechanism	√	-	√	√	√	√	√	[37]
Health and social awareness instillation	√	-	-	√	√	√	√	[6, 14]
Vocational education for human resource preparation*	-	-	-	√	-	-	-	-
Storage point	√	-	-	√	-	√	-	[10, 37]
Professional management of secondary market	-	-	√	-	-	-	√	[26]
Business localization*	-	-	-	-	-	√	-	-
Tax arrangement (additional or reduction) *	√	-	-	-	-	-	-	-
Jobs creation	√	-	-	-	√	-	√	[37]
Knowledge building and deployment	√	-	√	√	√	√	-	[20]
Program of restructuration machine*	-	√	-	-	-	-	-	-
Budget provision from APBN*	√	√	-	√	-	√	√	-
Regulation alignment between departments*	√	-	√	√	√	√	-	-
New entrepreneur invention program	-	√	-	-	√	-	√	[14, 36, 38]
Third party provision	√	√	-	-	-	√	-	[8, 20, 9, 37, 44]
Social Punishment*	√	-	-	-	-	√	-	-
Synergy of key actors	√	√	√	√	√	√	√	[34, 37]
Integrated implementation	√	√	-	√	√	√	√	[34, 36]
Control mechanism of RL implementation*	√	-	-	-	-	-	-	-
Green industrial development	-	√	-	√	-	√	-	[36]
SMEs-Industry link and match*	-	√	-	-	√	-	√	-
Process modernization of SMEs*	-	-	-	-	√	-	√	-
Tariff adjustment	-	-	-	-	-	-	-	[8, 10]
Waste to energy program	√	√	-	-	-	√	-	[8, 10]
Social and public facility utilization*	-	-	-	-	-	-	√	-
Monopolistic collective systems	-	-	-	-	-	√	-	[42]
All Stakeholders:								
Environmental preservation	√	√	√	√	√	√	√	[3, 6, 8, 10, 26, 34]
Prestige consideration (image, reputation)	√	√	-	√	-	√	√	[6, 14, 36, 41]

After sales service is needed to extend the life of electronic products. Usually, damaged electronic products are repaired for reuse. If those products cannot be used anymore, the owner will trade in with the new one in a secondary market. Unfortunately, usually retailers are only willing to buy these products at the lowest price, so that, the owners often cancel to sell it and prefer to keep it at home.

There are 7 actions recommended by government for community and only one action was not found in literature, i.e. fairness in using electronic product. Finally, this research found 28 actions that must be done by government itself. These phenomena give an impression that governments in many emerging countries have not paid great attention to SRL implementation, so they get the most responsibility than others.

Table 9: Linkage of reverse logistics and sustainability concept

Sustainable Reverse Logistics Categories	The Reasons to Conduct SRL			
	Profit	People	Planet	Reputation
Businesses:				
Eco-friendly product design		√	√	√
After sales service		√		√
Packaging reduction	√		√	√
Deposit fund	√			
Take back obligation			√	√
Drop box facility		√	√	
Recycling technology creation			√	√
Standard equipment utilization		√	√	
Skill accreditation (employee competency)		√		
Quality of product (extension of longevity)	√			√
Organizational commitment				√
Budget allocation	√			
Inclusion of post-use product handling		√	√	
Technology conversion			√	√
SMEs coaching		√		
Enterprise alliances	√		√	
Inclusion of product content and duration		√		
Corporate citizenship		√	√	√
Public:				
Public participation awareness			√	√
Eco shopping decision	√		√	√
Reduce replacement frequency			√	√
Separate own e-waste		√	√	√
Reasonable reimbursement (incentives)	√			
Willingness to pay recycling cost			√	√
Fairness in using electronic product			√	√
Government:				
Illegal import restriction			√	
Elect. prod. national stand. arrangement		√	√	√
Collecting point extension			√	
Invention of handling mechanism			√	
Health and social awareness instillation		√		
Vocational education for HR preparation		√	√	
Storage point			√	
Professional mgt. of secondary market	√	√		
Business localization		√	√	
Tax arrangement	√			
Jobs creation		√		
Knowledge building and deployment		√		
Program of restructuration machine		√		
Budget provision from APBN	√			
Regulation alignment between departments			√	
New entrepreneur invention program		√		√
Third party provision			√	
Social Punishment		√		
Synergy of key actors		√	√	
Integrated implementation			√	√
Control RL implementation			√	
Green industrial development			√	√
SMEs-Industry link and match		√	√	
Process modernization of SMEs		√	√	
Tariff adjustment	√			
Waste to energy program	√	√	√	√
Social and public facility utilization		√		
Monopolistic collective systems		√	√	√

To ensure that RL activities are very closely with sustainability concepts, each construct is grouped into four elements of sustainability depicted in Table 9. Interaction between RL categories and each element in sustainability concepts reflects the reasons why the stakeholders conduct SRL activities. However, government has never taken profit for government itself. The benefits gained from SRL activities are used for community welfare in form of circular economy.

Based on literature review and interview results, the duties of businesses suggested by government encompass creating eco-friendly electronic product design for easy features upgrading, providing after sales service to avoid quick replacement, reducing product package to save resources, initiating deposit fund to pay the environmental recovery, obligating take back program to provide collecting facilities, building drop box to facilitate community disposing used electronic product, finding recycling and e-waste conversion technology, utilizing standard equipment for maintaining society health, improving employee competency in e-waste processing. The business should strive to create an extension longevity of product, increase the organization commitment to implement RL especially in budget allocation, train the SMEs to process e-waste, publish how to handle product after using and product content, and overall, grow corporate citizenship. Government expects society to aware and participate in recycling program, by separating their own e-waste, eco-shopping, and reducing electronic product replacement. Government also hopes the public willing to pay cost of environmental recovery. Using electronic product fairly is an assertion that cannot be avoided. Although communities have also an obligation to protect the environment, they still expect to receive the reasonable reimbursement of their used electronic product unless they are reluctant to give their product.

The first government action to ensure SRL program will be implemented properly is restriction of illegal import used electronic products, since those products and their parts often cannot be utilized and only littering the environment. Moreover, informal businesses that seeks to profit from those products often ignoring safety and health process in reverse chain. Therefore, government should immediately make a regulation and national standard of import electronic product that will be sold in Indonesia and how to treat them, so they are not quickly pollute the environment.

This research finds some communities' expectations from a preliminary research, for example, collecting point facility to get a lot of amount of e-waste should be provided by government, because e-waste is a resource for SRL activities [4]. They should also easy to find a drop box by providing it close to the community. While based on businesses perspectives, government should motivate communities to give their used electronic product by giving a magnitude incentive. Regarding incentive for communities, the researcher suggests to government to make an incentive mechanism or regulation The Environment and Hygiene Office has a pick-up program for used electronic products, however, they do not give incentive to owners so they are often unmotivated to hand over those products for recycling. So that, instillation of health and social awareness in conducting SRL activities are also a big homework for government in emerging countries.

4.5. The Sound of Sustainable Reverse Logistics Theory building for Emerging Countries

Building SRL theory particularly for emerging countries is not an easy problem. Government should start from providing a good infrastructure including technology, setting up human resource capacity including their awareness and knowledge maturity, formulating the regulation and policies, arranging the mechanism, and designing monopolistic collective system. If the process mechanism in reverse chain has been running, the next step is providing professional management of secondary market. According to the head of sub directorate of e-waste and recycle, for implementing SRL program, Indonesian Government must depart from making a policy. After that the policy should be socialized so community wants to participate. While communicating the program, government must immediately build the infrastructure, and the last step is program enforcement at all level of society, if needed, mechanism enforcing of punishment should be prepared. Integrated implementation and all stakeholder involvement are also needed to ensure that all RL activities are in line with quattro bottom line concept.

This research is categorizing the stakeholder into two groups, i.e. internal and external stakeholder. Internal stakeholder consists of consumer, manufacturer, and intermediary business. While, external stakeholder comprises of environment, society, and government itself. Government intervenes other stakeholders to support and conduct SRL program with regard to the principle of sustainability in order not to harm other interests (Figure 3).

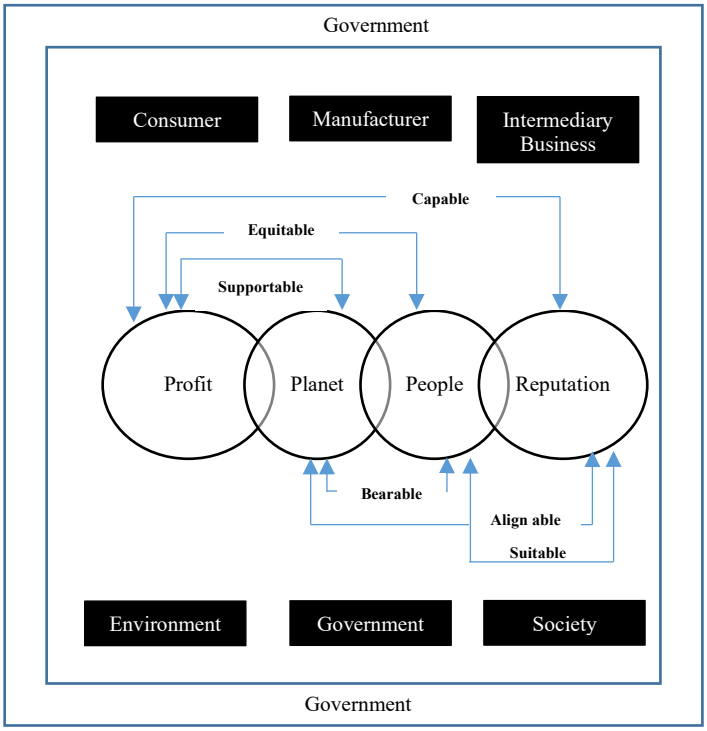


Figure 3: Quattro Bottom Line Approach for Sustainable Reverse Logistics Theory

The next step in building SRL theory is rearranging individual practices and literatures into "meta" constructs. In general literature review that had low levels of adoption, limited applicability and/or which did not relate to other practices were dropped so for further analysis we only use 51 constructs. The SRL theory to be built refers to two themes, namely activities related to RL implementation that must be conducted by all stakeholders based on government perspective and sustainability concepts with its four elements, i.e. profit, planet, people, and reputation (Figure 3). In building the SRL theory, the researcher seeks to balance between two interests. Therefore, this research formulates the sixth balancing constructs, namely: supportable, equitable, capable, bearable, align able, and suitable that will be described below,

- Supportable, re-design the way our economy works and re-build a restorative economy through creativity and also innovation, generate business and economic opportunity (Profit) and provide environmental benefits through renewable and reusable resources as materials in an efficient way (Planet).
- Equitable, conformity between opportunities to create greater value and align incentives (Profit) and societal benefits through invention of job creation and additional income in reverse chain (People)
- Capable, ability to maximize resources in-use, maintain, repair, and upgrade the lifetime of electronic products will bring benefit (Profit) and demonstrate positive image (Reputation).
- Bearable, alignment of responsibility to maintain the environmental preservation (Planet) and global healthcare activity in reverse chain (People)
- Align able, fair phone design in using materials, good working conditions to enable the reuse and recycling of parts (Planet) will anchor deep niche position in the market (Reputation)
- Suitable, worthy activities in supporting the program to reach social welfare (People) and to be responsible in using electronic product and/or handling their own e-waste (Reputation).

Ideally a theory is built by adjusting why the theory is needed, where that theory will be applied, who the actor is, and what the outcome and expectation are. As explained in previous section, the SRL theory is needed for handling e-waste problem due to low of public awareness in using electronic product especially MP and almost of them do not know how to treat the used MP that is no longer used. Therefore, the SRL theory that will be applied in emerging countries, has to be made and initiated by government. The theory is designed to encourage other stakeholders conduct RL practices accordance with expected output and outcomes. After describing the constructs, either from literature or individual practice (Table 8 and 9) and exploring the possible interaction between sustainable constructs (Figure 3), this research categorizes these constructs to adopt parsimony and complete requirement in theory building. As explained in this research method section, this research uses three types of evaluations, namely criterion-focused evaluation, impact focus, and implementation focus. Table 10 is the end result of this analysis. Due to this theory building is built from government perspective, this research has to consider the overall vision, mission, function, and primary task of government. This research summarized those considerations into four categories, namely infrastructure and facilities provision, human resource commitment, regulation, and managerial orientation.

Table 10: Rearranging literatures review and individual practices into meta constructs

Meta Constructs	Literatures Review and Individual Practices	Number of Government Recommendation
Infrastructures and Facilities	1. Converting technology creation	6
	2. Government budget provision	5
	3. Drop box facilities	4
	4. Collecting point extension	4
	5. Additional storage point	3
	6. Job creation	3
	7. Social and public facility utilization	1
Human Resource Commitments (behavior and attitude)	1. Standard equipment utilization	5
	2. Health and social awareness instillation	5
	3. Public participation awareness	5
	4. Reduce replacement frequency of electronic product	3
	5. Individual commitment	3
	6. Eco-shopping decision	2
	7. Willingness to pay recycling cost	2
	8. Fairness in using electronic products	2
	9. Corporate citizenship	-
Regulations	1. Synergy of key actors	7
	2. E-waste handling mechanism	6
	3. Integrated implementation	6
	4. Skill accreditation	6
	5. Regulation alignment between departments	5
	6. Building knowledge and deployment	5
	7. Coaching for SMEs	5
	8. Separate own e-waste	3
	9. Take back obligation	3
	10. Waste to energy program	3
	11. SMEs-Industry link and match	3
	12. Third party recycling business provision	3
	13. New entrepreneur invention program	3
	14. Enterprise alliance	2
	15. Inclusion of product content	2
	16. Inclusion of post-use product handling	2
	17. Reasonable reimbursement	2
	18. Illegal import restriction	2
	19. Social punishment	2
	20. Professional management of secondary market	2
	21. Preparation vocational education to increase individual capability	1
	22. Localization of reverse logistics business	1
	23. Tax arrangement both additional for new product or reduction for second hand product	1
	24. Program of restructuration machine	1
	25. Control mechanism of RL implementation	1
	26. Process of SMEs modernization	1
	27.Monopolistic collective system	1
	28. Deposit fund	1
Managerial Orientation, Vision and Mission	1. Budget allocation	6
	2. Product quality	5
	3. National standard arrangement particularly for electronic product	3
	4. Green industrial orientation	3
	5. After sales service	3
	6. Eco-friendly product design	2
	7. Packaging reduction	1

Category 1 – Infrastructure and Facilities. There are 7 selected constructs in this category. Converting technology creation is the most recommendation from government. It was recommended by 6 national agencies. if an emerging country has this technology, a lot of e-waste will not be a problem instead a resource. For providing it, government should prepare a large budget. Moreover, the budget is not only for technology procurement, but also to add drop box facilities, and extend the collecting and storage point.

Although Ministry of Environment and Forestry has the largest responsibility, however, RL cannot be implemented without supporting from all parties include from other departments. Two of departments did not have good alignment between the element of sustainability, especially between profit and other elements. Some efforts to reduce plastics usage will never work due to other departments (Ministry of Industry and Ministry of trade) encourage precisely to produce it for the reason gaining economics and industrial growth.

Actually, Indonesia has a regulation to ban importing waste and used electronic product from other countries, however, we often still found those products in many places such as Wakatobi, Batam, Surabaya, Pare-Pare, and even in the capital city of Indonesia, Jakarta. There is no control mechanism to prevent importing e-waste, integrated RL implementation for handling e-waste problems, and significant investment for building a big industry in reverse chain.

Category 2 – Human Resource Commitment. This research finds 9 constructs in category of human resource commitment. This category describes communities' (people) attitude toward SRL program. Public awareness is the central issue in this category for conducting SRL program. We divided public awareness into three dimensions, namely 1) awareness to use basic standard equipment such as rubber gloves, masker, and so on include taking care of community health; 2) awareness to actively participate in recycling program such as separating their own e-waste and willing to pay the recycling cost; and 3) awareness to prevent the increasing amount of e-waste by reducing replacement frequency, eco-shopping, and fairness in using electronic products. That e-waste problem is a small matter that does not need to get priority attention, e-waste is invaluable product and not as energy resources, e-waste problem is not my backyard and other negative syndrome have to be eliminated from societies' mind. Not only societies that should be aware, but also businesses and government itself must be conscious and should be the first guard of environment protecting initiative. Environment and Hygiene Office Jakarta Province failed to put the drop box facilities in each Transjakarta bus stop until now for a reason they do not know. Commitment to successfully implementing activities in reverse chain were evidenced in seventh ways [28], i.e.: 1) credibility of decision maker, in this case is government; 2) implementation mechanism include incentive to market; 3) significant attention of secondary market; 4) societies involvement; 5) sustainable commitment to high quality of operation facilities for managing e-waste; 6) Effectiveness evaluation to the strategy chosen; and 7) the clear legal protection for handling e-waste by government.

Category 3 – Regulation. This research found a lot of government responsibilities to regulate SRL program. The researcher believes these regulations (categories) will be the measurement of government performance in SRL implementation.

In the first type evaluation, this research found five evaluation criteria of SRL implementation. The evaluation criteria are classified into five dimensions, namely integrated enforcement of SRL program by all parties even among government agencies, obligation compliance to support SRL implementation, competency improvement to prevent SRL activities polluting the environment, creative ideas development to create income in reverse chain, and professional management to arrange SRL implementation properly.

This research establishes synergy of key actors such as government, businesses, and community; integrated SRL implementation; regulation alignment between department; and enterprise alliances as government responsibilities in integration enforcement of SRL program. While, obligation compliance dimension has seven indicators, namely willingness to separate e-waste by sorting the valuable material, conduct take back scheme, publish product content and post-use product handling in packaging, arrange reasonable reimbursement system, prevent illegal import of used electronic products, and socialize restructuration machine program to everyone that needs recycling machine but does not have enough money to buy it. Improvement of human resource competency is very important for implementing SRL, because to process used electronic product required a specific skill. Therefore, government responsible to increase return handling capability by giving a training and an accreditation certificate for individual or firm that really has recycling ability. Big companies are usually obliged to train small medium enterprises to transfer their skill in recycling process. Many developed countries such as Sweden and Japan provide vocational education specifically for increasing the skill in reverse chain, so they can operate the RL activities without damaging the environment and community interests.

The government responsibilities in creative idea development criteria comprise five indicators, namely building waste to energy (WTE) program, providing third party recycling business, inventing new entrepreneur in reverse chain field, organizing monopolistic collective system, and constructing deposit fund scheme. Many emerging countries have not been able to take benefit from e-waste, whereas developed countries import a lot of e-waste from their neighboring countries to be converted into energy. WTE needs a lot of resources, so the government in developed countries organizes monopolistic collective system to avoid informal group intervention. Nevertheless, Indonesian government is always trying to find new entrepreneurs in reverse chain field, due to many people in emerging countries make this activity as their primary livelihood. For businesses that have not been able to recycle their used product or to neutralize their hazardous products, government should provide third party recycling business. Since converting used electronic product into energy or other valuable products needs a lot of funds, constructing deposit fund scheme is urgent to do immediately.

The last regulation for SRL implementation is professional management that has the most number of indicators, i.e. professional in e-waste handling mechanism to avoid too large intervention of informal group, SMEs-industry link and match to improve employees' skill, social punishment application for tight SRL implementation in avoiding environmental pollution, secondary market management to develop business in reverse chain, business localization in reverse chain for easy controlling the SRL activities, SMEs modernization, and tax arrangement to grow the business spirit.

Category 4 – Managerial Orientation. In the forth category, this research found seven dimensions. The first government duty is providing budget for the expensive technology, facilities, and infrastructures and guarantee the quality of import electronic products that will be sold in Indonesia to prevent those products littering the environment quickly. Today, Indonesian government does not have a national standard for electronic products, so there are many unqualified products sold in Indonesian market. Besides easily damaged, these products are harmful to human body since they contain toxic and hazardous materials.

In managerial orientation, vision, and mission, government also should think about how to manage SRL activities from collecting process until selling the products to secondary market by using environmental friendly ways through reducing the use of packaging and trying to create eco-friendly products. Besides that, after sales service believed can reduce e-waste generation since this action adds life span of the products. Actually, all government agencies (respondents) recognize that environmental preservation is their responsibility. They are conscious that to protect the environment they should cooperate with others [2].

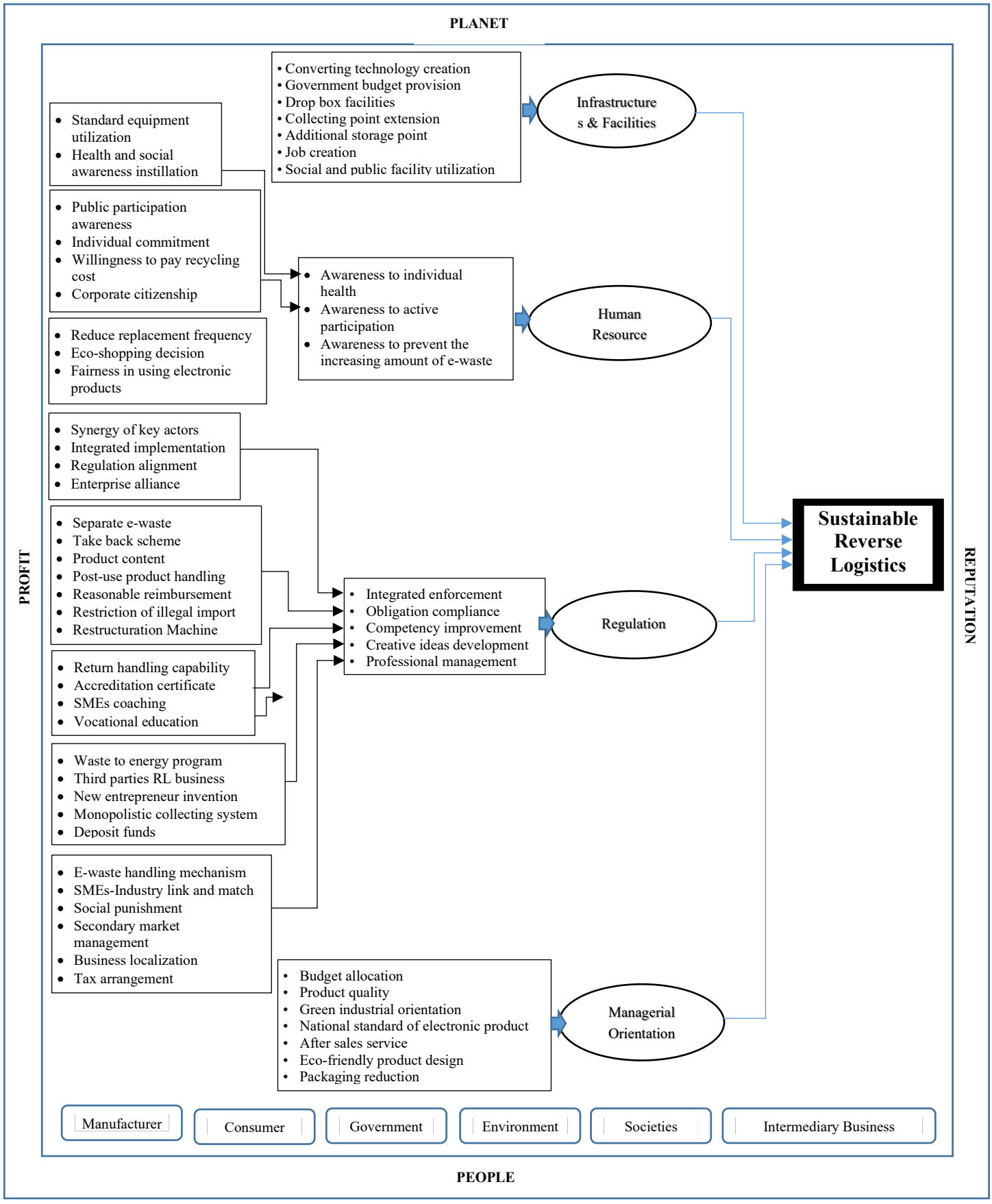


Figure 4: A model of sustainable reverse logistics for emerging countries from government perspectives

They know also that community in emerging countries usually have low awareness to protect the environment, therefore, the initiative must come from the government. Unfortunately, in many emerging countries, the government itself has often not felt that activity in reverse chain currently is very important due to the abundant of production in forward chain. Some respondents, especially Ministry of Environment and Forestry and Environment and Hygiene Office plans to apply punishment to arouse community awareness. They believe that must use the hard way. Besides that, they have to think how to implement and control that regulation.

After finding the evaluation criteria, this research uses impact evaluation for describing the direct and indirect impact of the use of indicators, dimensions, and variables formatted to measure the successful of SRL theory implementation, that is a reverse logistics theory that constantly in line with four elements in sustainability concepts, i.e. profit, planet, people, and reputation. SRL theory is recommended if activities in reverse chain will be applied in emerging countries. Using reverse logistics terminology is perceived to be no longer appropriate since many people ignore environmental friendly way. Therefore, sustainable reverse logistics concept is a theory which should be standardized for emerging countries.

Figure 4 is the last figure and the end of finding in this research that summarize all constructs for building SRL theory from government perspectives. Quattro bottom line as an approach that can preserve RL activities will not interfere other interests and the sixth ability to keep the balancing among interests and stakeholders will consider to be used in this research.

This research uses implementing-focused evaluation in the last step to test the theory building. For this type of evaluation, this research has two main propositions and eighteen proponent propositions to build a whole theory of sustainability reverse logistics and explain each interaction between two spheres, i.e.: capable, equitable, align able, supportable, suitable and bearable. However, this research does not examine the theory, but only formulate the propositions.

Proposition 1: The infrastructures and facilities provision, human resource commitment, regulation establishment, and managerial orientation are required to build a sustainable reverse logistics.

Proposition 1a: Converting technology creation, budget provision by government, providing drop box facilities, collecting point extension, additional storage point, job creation, and social and public facility utilization are needed for infrastructures and facilities provision.

Proposition 1b: Awareness to individual health, active participation, and prevent the increasing amount of e-waste are needed to build human resource commitment.

Proposition 1b1: Utilization of standard equipment and instillation of health and social awareness are needed to build individual health awareness

Proposition 1b2: Public participation, individual commitment, willingness to pay recycling cost, and corporate citizenship are believed can build awareness to active participation.

Proposition 1b3: Reduce replacement frequency, eco-shopping decision, and fairness in using electronic product are needed to prevent the increasing amount of e-waste.

Proposition 1c: Integrated enforcement, obligation compliance, competency improvement, creative idea development, and professional management should be regulated immediately.

Proposition 1c1: Synergy of key actors, integrated implementation, alignment of regulation, and enterprise alliance are needed to integrate the implementation.

Proposition 1c2: Separating own e-waste, taking back scheme, inclusion of product content and post-use product handling, reasonable reimbursement, restriction of illegal import, and restructuration machine are the form of obligation compliance.

Proposition 1c3: Return handling capability, accreditation certificate, SMEs coaching, and vocational education are needed to improve the competency.

Preposition 1c4: Waste to energy program, third parties recycling business, new entrepreneur invention, monopolistic collecting system, and deposit funs are the creative ideas that should be regulated immediately.

Preposition 1c5: E-waste handling mechanism, link and match between industry and SME's, social punishment, management of secondary market, business localization, tax arrangement, and modernization of SMEs need professional management.

Preposition 1d: Budget allocation, product quality, green industry, national standard for electronic products, after sales service, eco-friendly product design, reduce packaging are the types of managerial orientation.

Preposition 2: Sustainable reverse logistics could be realized by referring quattro bottom line conceptual framework.

Proposition 2a (capable): Stakeholders capability to balance target profit achievement and reputation recognition is required to create sustainable reverse logistics.

Proposition 2b (equitable): Stakeholders fairness between obtaining profit and preserving social harmony is required to realize sustainable reverse logistics.

Proposition 2c (align able): Align able of Stakeholders interest between to reach reputation recognition and environmental conservation is required to establish sustainable reverse logistics.

Proposition 2d (bearable): Stakeholders responsibility to preserve social harmony and environmental conservation is required to create sustainable reverse logistics.

Preposition 2e (suitable): worthy activities in supporting the program to reach social welfare and to be responsible in using and producing electronic product and/or handling their own e-waste are the suitable components to format sustainable reverse logistics

Preposition 2f (supportable): re-design the way our economy works and re-build a restorative economy through creativity and also innovation, generate business and economic opportunity and provide environmental benefits through renewable and reusable resources as materials in an efficient way.

5. Conclusion and Further Research

This research uses criteria of good theory [29, 32] to break down few conclusions. First, this research highlights 51 selected constructs for building sustainable reverse logistics theory that categorized into 4 dimensions to fulfil parsimony requirement and evaluation criteria in building a theory. They are infrastructures and facilities provision, human resource commitment, regulation, and managerial orientation. Second, the causality relationship of major constructs both direct and indirect impact of SRL implementation, is used to explain the internal consistency and impact criteria. Third, the implementation criteria are met by finding the stakeholders obligations to realize SRL program. There is a small chance for community or businesses in emerging countries to initiate conduct SRL activities, therefore these constructs are the obligations for them. In the case of Malaysia, manufacturers are not proactive in taking advantage of product take back since they regard it as service cost [36, 46]. While in Thailand, they still face many problems in conducting RL such as return policy, procedure management, transportation and communication [6]. The author believes that many poor and emerging countries still grapple with the same problem. Fourth, in term of fulfilling generalization criterion, it is better to test in some emerging countries with the same phenomena before using this theory. For further research, it is needed to explore the new constructs based on businesses and communities' inspirations. Furthermore, for managerial implication, government will adopt those constructs to measure SRL performance.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study would not have been possible without support material and non-material from LPDP scholarship from The Ministry of Finance and The Ministry of Research Technology and Higher Education. The author would like to thank and appreciate the invaluable constructive views and comments by reviewers.

References

[1] H. Maheswari, G. Yudoko, A. Adhiutama, "Towards sustainable reverse logistics: A conceptual framework of quattro bottom line approach" in Indstr. Eng. and Eng. Mgt Confr. 2017, 1377-1381,

[2] G. Yudoko, "Integrated municipal solid waste planning and management (IMSWPM) in developing countries: A conceptual framework", J. Mgt. Tech., 2007

[3] Organization for Economic Cooperation and Development (OECD), "OECD economic survey: Indonesia", OECD Publishing, 2010

[4] H. Maheswari, G. Yudoko, A. Adhiutama, "Stakeholder engagement model in quattro helix model for mobile phone reverse logistics in Indonesia: A conceptual framework" in IOP Conf. Series: Mat. Sci. and Eng., 2017. https://doi:10.1088/1757-899X/277/1/012062

[5] P. Andarani, N. Goto, "Potential e-waste generated from households in Indonesia using material flow analysis" J Mater Cyc. Waste Mgt., **16**, 306-320, 2014. http://doi.org/10/1007/s10163-013-0191-0

[6] P. Tanskanen, Electronics waste: recycling of mobile phones, INTECH Open Access Publisher, 2012

[7] J.D. Demajorovic, E.E.F. Augusto, M.T.S. De Souza, "Reverse logistics of e-waste in developing countries: challenges and prospects for the Brazilian model", Ambt. And Soc, **19** (2), 117-136, 2016

[8] P.R. Kleindofer, K. Singhal, L.N.V Wassenhove, "Sustainable operations management" Prod. and Oprt. Mgt., **14** (4), 482-492, 2005. ABI/INFORM Global

[9] P.J. Daugherty, A.E. Ellinger, "Reverse Logistics: The relationship between resorce commitment and program performance", J. of Busn. Logst., **22** (1), 107-123, 2001

[10] L. Paquette, "Reverse logistics and the creations of reverse flows for the mobile phone industry" Thesis, Aarush University, 2009

[11] D.S. Rogers, P. Tibben-Lembke, "Going Backwards: Reverse Logistics Trends and Practices", Reverse Logistics Executive Council, University of Nevada, Reno Center for Logistics Management, 1998

[12] H.C. Pfohl, A. Bode, H.T.V. Nguyen, "Adaptability to reverse logistics – An empirical study in European electronics industry", Journal of Glob. Strat. Mgt., **6** (1), 89-102, 2012. http://doi.org/10.20460/JGSM.2012615789.

[13] I. El-Nakib, "Reverse logistics: A comparison of electronic waste recycling between Switzerland and Egypt" in Global Conference on Operations and Supply Chain Management, 2012, ISBN No. 978-967-5705-06-9

[14] M.D. Abdulrahman, N. Subramanian, "Barriers in implementing reverse logistics in Chines Manufacturing Sectors: An Empirical Analysis" in the 23rd Annual POMS Conference, 2012

[15] A. Agrawal, "Reverse logistics: Performance measure and their effect in product lifecycle" in Inter. Jour. of Core Eng. and Mgt, **1** (2) 14-23, 2014, ISSN: 2348 9510

[16] S. Emmet, V. Sood, Green Supply Chain, A John Wiley and Sons, Ltd. Publication, United Kingdom, 2010

[17] E. Damanhudi, "The role of informal collector of recyclable waste and used goods in Indonesia", Post-Cons. Waste Rec. and Opt. Prod., Intechopen.com, 2012, Croatia

[18] R.F. dos Santos, F.A.S. Marins, "Integrated model for reverse logistics management of electronic products and components" Procedia Computer Science, **55**, 575-585, 2015

[19] M. Thierry, M.J.V.N. Salomon, J.V. Nunen, L.V. Wassenhove, "Strategic issue in product recovery management", California Mgt. Rev., **37** (2), 114-135, 1995, http://doi.org/10.2307/41165792

[20] M.P. De Brito, R. Dekker, "A framework for reverse logistics", ERIM Rep. Series, 1-23, 2003, http://doi.org/10.1007/978-3-540-24803-3_1

[21] S.S. Khan, S.A. Lodhi, F. Akhtar, "Sustainable WEEE management solution for developing countries applying human activity system modeling", Mgt. of Envr. Qual., **26** (1), 84-101, 2014, http://doi.org/10.1108/MEQ-05-2014-0072

[22] V.P. Rindova, I.O. Williamson, A.P. Petkova, J.M. Sever, "Being good or being known: An empirical examination of the dimensions, antecedents, and consequences of organization reputation", The Acadm. Of Mgt Journ, **48** (6), 1033-1049

[23] C.E. Carroll, "Corporate Reputation and the News Media in the United States" New York, NY: Routledge, 2011

[24] T. Passow, R. Fehlmann, H. Grahlow, "Country reputation – from measurement to management: The case of Liechtenstein, Corport. Rept. Rev, **7** (4), 309-326, 2005, http://doi.org/10.1057/palgrave.crr.1540229

[25] A. McKinnon, The role of government in promoting green logistics", The Chart. Inst. of Log. and Transp., 342-360, ISBN 97807494567888, 2010, United State.

[26] J. Aitken, A. Harrison, "Supply governance structures for reverse logistics systems", Intr. J of prt. and Prod. Mgt., **33** (6), 745-764, 2012, DOI: 10.1108/IJOPM-10-2011-0362

[27] W. Wilhelm, A. Yankov, P. Magee, "Mobile phone consumption behavior and the need for sustainability innovation", J. of Strt. Innvt. and Sust., **7** [2] 20-38, 2011

[28] M.A.P. Sari and Rustan, Implementation of good governance in waste management", PKP2A III LAN Samarinda (2013)

[29] R. Swanson, T.J. Chermack, Theory Building in Applied Disciplines", Berrett-Koehler Publ., 2013, San Fransisco

[30] D.K. Beckley, W.B. Logan, "The Retail Sales Person at Work", Gregg Publs., New York, NY, 1948

[31] M.Q. Patton, "Utilization-Focused Evaluation: The New Century Text", Sage Publicatio, UK, 1997

[32] J.G. Wacker, "A definition of theory: Research guidelines for different theory building research method in operation management", J. Opr. Mgt., **16,** 361-385, 1998

[33] M. Bernon, J. Cullen, "An integrated approach to managing reverse logistics", Intrn. J. of Log. Resrc. and Applc., **10**, 41-56, 2007

[34] C.H. Chiou, "Dynamic capabilities, collaborative network and business model: An empirical analysis of Taiwan HTC corporation", Affrican J. of Busn. Mgt. **5** (2), 299-305, 2011

[35] H.L Correa, L.H. Xavier, "Concepts, design, and implementation of reverse logistics systems for sustainability supply chain in Brazil", J. of Oprt. And Supp. Cha. Mgt., **6** (1), 1-25, 2013

[36] S.M. Budijati, Model of Reverse Logistics Management: Close Loop and Opened Loop as well as Consumer Behavior Accommodation, Dissertation, UGM, 2016

[37] J-B. Sheu, "Green supply chain collaboration for fashionable consumer electronics products under third-party power intervention-a resources dependence perspective" Sustainability, **6**, 2832-2875, 2014. http://doi.org/10.3390/su6052832

[38] R.G. de Sauza, J.C.N. Climaco, A.P. Sant'Anna, T.B. Rocha, R.A.B. Valle, O.L.G. Quelhas, "Sustainability assessment and prioritization of e-waste management", Wst. Mgt., **57**, 46-56, 2016, http://dx.doi.org/10.1016/j.wasman.2016.01.034

[39] I. Vlachos, "A conceptual framework of reverse logistics impact on firm performance", Brit. Acdm. of Mgt (BAM) Confr., 2014

[40] P. Rajagopal, V.P.K. Sundram, B.M. Naidu, "Future direction of reverse logistics in gaining competitive advantage: A review literature", Int. J Sup. Chain. Mgt, **4** (1), 39-48, 2015

[41] J. Yin, Y. Gao, H. Xu, "Survey and analysis of consumers' behavior of waste mobile phone recycling in China", J of Clean. Prod., 1-9, 2013, http://dx.doi.org/10.1016/j.jclepro.2013.10.006

[42] R. Pinna, P.P. Carrus, "Reverse logistics and the role of fourth party logistics provider", Pathw. to Sup. Ch. Exc., www.intechopen.com

[43] A. Atasu, L.V.N. Wassenhove, "An operation perspective on product take-back legislation for e-waste: Theory, practice, and research need", Prod and Opr. Mgt., **21**, (3), 407-422, 2012, ABI/INFORM Research

[44] H. Maheswari, "Supply Chain Management", Cent. Of Teach. Mat. Dev., Economic Faculty, Mercu Buana University, 2008.

[45] K.S. Khor, Z.M. Udin, "Impact of reverse logistics product disposition towards business performance in Malaysian E&E companies", J. of Sup. Ch. And Cust. Rel. Mgt., 19 pg., 2012, DOI: 10.5171/2012.69946

Appendix

Constructs	Definition
Eco-friendly product design	Designing an electronic product that using eco-friendly resources and easy to upgrade the feature
After sales service	Giving service after selling electronic product
Packaging reduction	Reducing the use of packaging by alternating usage
Deposit fund	Providing budget by collecting from community for normalizing environmental condition
Take back obligation	Designing return electronic product mechanism from point of consumption (consumer) to point of origin (producer)
Drop box facility	Providing drop box to facilitate community bring and collect used electronic product
Recycling technology creation	Creating a technology to recycle used electronic product
Standard equipment utilization	Enforcing employee to use standard equipment when conducting RL activities
Skill accreditation (employee competency)	Accrediting the employee to standardize competency of employee
Quality of product (longevity of extension)	Establishing the quality of product by adding longevity of extension
Organizational commitment	Preparing commitment of organization to conduct RL activities
Budget allocation	Allocating budget to convert e-waste to be a valuable product
Inclusion of post-use product handling	Publishing how to handle after using electronic product
Technology conversion	Inventing a technology to convert e-waste to energy
SMEs coaching	Coaching small medium enterprise to recycle safely used electronic product
Enterprise alliances	Alliancing enterprises for easy sharing the knowledge and facility
Inclusion of product content and duration	Writing electronic product content and duration in guidance book for informing the when they allow to replace it
Corporate citizenship	A set of values or principles that drive an organization to become responsible engaged with particular RL activities [36]
Public participation awareness	
Eco shopping decision	Warning community to be aware to participate in RL program
Reduce replacement frequency	Buying electronic product when it is needed
Separate own e-waste	Reducing the frequency of electronic product replacement
Reasonable reimbursement (incentives)	Separating communities' own e-waste from household waste so that easy to recycle
Willingness to pay	Giving reasonable incentive to motivate owner of used electronic product to bring it for recycling
Fairness in using electronic product	Influencing community want to pay e-waste converting cost
	Using electronic product with reasonable way
Illegal import restriction	
Electronic product national standard arrangement	Preventing import of electronic product illegally
Collecting point extension	Making national standard for electronic product
Invention of handling mechanism	Adding collecting point for electronic products in many place
Health and social awareness instillation	Inventing mechanism of e-waste handling for each stakeholder
Vocational education for HR preparation	Bringing strong understanding of health and social awareness
Storage point	Developing vocational education to teach safety recycling technology
Professional mgt. of secondary market	Building storage point to relocate used electronic products from drop box
Business localization	Managing secondary market professionally to facilitate circulation of used electronic product
Tax arrangement	Placing businesses in reverse chain on particular area
Jobs creation	Arranging tax reduction for reselling used electronic products
Knowledge building and deployment	Creating a new job in reverse chain to help uneducated people as part of recycling program e.g. as scavenger.
Program of restructuration machine	Building knowledge to community about waste, hazardous and toxic materials
Budget provision from APBN	Subsidizing businesses by giving a new machine that is needed for converting e-waste to energy
Rules alignment	Providing budget to run RL program
New entrepreneur invention program	Aligning rules and regulation in each department for integrating RL program
Third party provision	Inventing new entrepreneurs that do business in recycling used electronic product
Social Punishment	Providing a company that certified by government to conduct RL activity for firms that do not have recycling technology
	Giving punishment to communities that avoid to separate their own e-waste from other kinds of waste, wasteful in using electronic product
Synergy of key actors	Integrating key actors to implement RL activities together
Integrated implementation	Integrating RL activities to reduce e-waste faster
Control RL implementation	Controlling implementation of RL program to ensure not damaging the environment
Green industrial development	Developing green industrial by making a competition
SMEs-Industry link and match	Creating link and match between SMEs and industry
Process modernization of SMEs	Improving process modernization of SMEs in managing business in reverse chain
Tariff adjustment	Adjusting tariff for electronic products that have longer life time
Waste to energy program	Implementing waste to energy program
Social and public facility utilization	Using social and public facility to facilitate business in reverse chain
Monopolistic collective systems	Dominating collective system of used electronic product to avoid uncertain timing and quantity of returns

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 92-99 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Evaluating the effect of Locking on Multitenancy Isolation for Components of Cloud-hosted Services

Laud Charles Ochei[*,1], Christopher Ifeanyichukwu Ejiofor[2]

[1]*School of Computing and Digital Media, Robert Gordon University, United Kingdom*
[2]*Department of Computer Science, University of Port Harcourt, Nigeria, christopher.ejiofor@uniport.edu.ng*

ARTICLE INFO

Article history:
Received: 13 April, 2018
Accepted: 11 May, 2018
Online: 31 May, 2018

Keywords:
Multitenancy
Tenant Isolation
Locking
Cloud-hosted Service
Cloud Patterns
Software Process
Bug Tracking

ABSTRACT

Multitenancy isolation is a way of ensuring that the performance, stored data volume and access privileges required by one tenant and/or component does not affect other tenants and/or components. One of the conditions that can influence the varying degrees of isolation is when locking is enabled for a process or component that is being shared. Although the concept of locking has been extensively studied in database management, there is little or no research on how locking affects multitenancy isolation and its implications for optimizing the deployment of components of a cloud-hosted service in response to workload changes. This paper applies COMITRE (Component-based approach to Multitenancy Isolation through Request Re-routing) to evaluate the impact of enabling locking for a shared process or component of a cloud-hosted application. Results show that locking has a significant effect on the performance and resource consumption of tenants especially for operations that interact directly with the local file system of the platform used on the cloud infrastructure. We also present recommendations for achieving the required degree of multitenancy isolation when locking is enabled for three software processes: continuous integration, version control, and bug tracking.

1. Introduction

Multitenancy (that is, an architectural practice of using a single instance of a service to serve multiple tenants) is a notable feature in many cloud-hosted services. Multiple users are usually expected to access a shared functionality or resource and so there is need to ensure that processes and data associated with a particular tenant and/or component does not affect others [1]. We refer to this concept as *multitenancy isolation*. Multitenancy isolation is a way of ensuring that the performance, stored data volume ad access privileges required by one tenant and/or component does not affect other tenants and/or components [1][2].

There are different or varying degrees of multitenancy isolation. For example, a higher degree of isolation would be imposed on a component that cannot be shared due to strict regulations than for a component that can be shared with minimal reconfiguration. A high degree of isolation implies that there is little or no interference between tenants when they are accessing a shared functionality/process or component of a cloud-hosted service, and vice versa. We can achieve a high degree of isolation by duplicating a component (and its supporting resources) exclusively for one tenant.

One of the conditions that can influence the degree of isolation is when locking is enabled for the functionality/process or component that is being shared. Locking is a well-known concept used in database management to prevent data from being corrupted or invalidated when multiple users try to read or write to the database [3]. Any single user can only modify items in the database to which they have applied a lock that gives them exclusive access to the record until the lock is released. The concept of locking in database management is closely related to multitenancy isolation in the sense that both of them are used to prevent multiple users from performing conflicting operations on a shared process or component and can also be implemented at different or varying degrees. Despite this similarity, there is little or no research on how locking affects multitenancy isolation and its implications for optimizing the deployment for components of a cloud-hosted service in response to workload changes.

[*]Corresponding Author: Laud Charles Ochei, Robert Gordon University
E-mail : l.c.ochei@rgu.ac.uk

Motivated by this problem, this paper applies COMITRE (Component-based approach to Multitenancy Isolation through Request Re-routing) to evaluate the impact of enabling locking for a shared process or component of a cloud-hosted application. This paper addresses the following research question: "How can we evaluate the required degree of multitenancy isolation when locking is enabled on a shared process or component of a cloud-hosted service?" To the best of our knowledge, this study is the first to apply an approach for implementing the required degree of multitenancy isolation for a shared process or component of a cloud-hosted service when locking is enabled and to analyse its impact on the performance and resource consumption of tenants. In this study, we implemented multitenancy isolation based on three multitenancy patterns (i.e., shared component, tenant-isolated component, and dedicated component) to analyse the effect of the different degrees of isolation on performance and resource consumption of tenants when one of the tenants is exposed to high workload. The experiments were conducted using a cloud-hosted continuous integration system using Hudson as a case study deployed on a UEC private cloud. The results showed that when locking is enabled, it can have a significant effect on the performance and resource consumption of tenants especially for operations that interact directly with the local file system of the operating system or platform used on the cloud infrastructure.

The main contributions of the paper are:

1. Applying the COMITRE approach to empirically evaluate the required degree of multitenancy isolation for cloud-hosted software services when locking is enabled.
2. Presenting how locking is used in three different software processes (i.e., continuous integration, version control and bug tracking) to achieve multitenancy isolation, and its implication for optimal deployment of components.
3. Presenting recommendations and best practice guidelines for achieving multitenancy isolation when locking is enabled.

The rest of the paper is organised as follows: Section two discusses the relevance locking to multitenancy isolation for cloud-hosted services. Section three is the methodology, and Section four presents the results and discussion. The recommendations and limitations of the study are detailed in Section five and six respectively. Section seven concludes the paper with future work.

2. Relevance of Locking on Multitenancy Isolation for Cloud-Hosted Services

Multitenancy is an important cloud computing property where a single instance of an application is provided to multiple tenants, and so would have to be isolated from each other whenever there are workload changes. Just as multiple tenants can be isolated, multiple components being accessed by a tenant can also be isolated. We define "Multitenancy isolation" in this case as a way of ensuring that the required performance, stored data volume and access privileges of one component does not affect other components of a cloud-hosted application being accessed by tenants.

When a component of a cloud-hosted application receives a high workload and there is little or no possibility of a significant influence on other tenants, we say that there is a high degree of isolation and vice versa. The varying degrees of multitenancy isolation, can be captured in three main cloud deployment patterns: (i) dedicated component, where components cannot be shared, although a component can be associated with either one tenant/resource or group of tenants/resources; (ii) tenant-isolated component, where components can be shared by a tenant or resource instance and their isolation is guaranteed; and (iii)shared component, where components can be shared with a tenant or resource instance and are unaware of other components.

Assuming that there is a requirement for a high degree of isolation between components, then components have to be duplicated for each tenant which leads to high resource consumption and running cost. A low degree of isolation may also be required, in which case, it might reduce resource consumption, and running cost, but there is a possibility of interference when workload changes and the application does not scale well.

Most of the widely used Global Software Development processes like continuous integration (for example, Hudson), version control (for example, with Subversion) and bug tracking (for example, with Bugzilla) implement some form of locking whether at the database level or filesystem level. In continuous integration for instance, locking can be used to block builds with either upstream or downstream dependencies from starting if an upstream/downstream project is in the middle of a build or in the build queue. Again, locking operations are also used in version control systems (e.g., subversion) and bug tacking systems (e.g., bugzilla) [3] [4] [5].

There are several research work on multitenancy isolation such as [6], [7] and [8]. However, none of these works have focused on the effect of locking on multitenancy isolation for components of a cloud-hosted service.

3. Evaluation

In the following, we present the experimental setup and the case study we have used in this study.

3.1. Applying COMITRE to Implement Multitenant Isolation

We applied COMITRE to evaluate multitenancy Isolation in a Version Control system. Fig. 1 shows the structure of COMITRE. It captures the essential properties required for the successful implementation of multitenancy isolation, while leaving large degrees of freedom to cloud deployment architects depending on the required degree of isolation between tenants. The actual implementation of the COMITRE is anchored on shifting the task of routing a request from the server to a separate component (e.g., Java class or plugin) at the application level of the cloud-hosted

GSD tool. The full explanation of COMITRE plus the step-by-step procedure and the algorithm that implements it is given in [9].

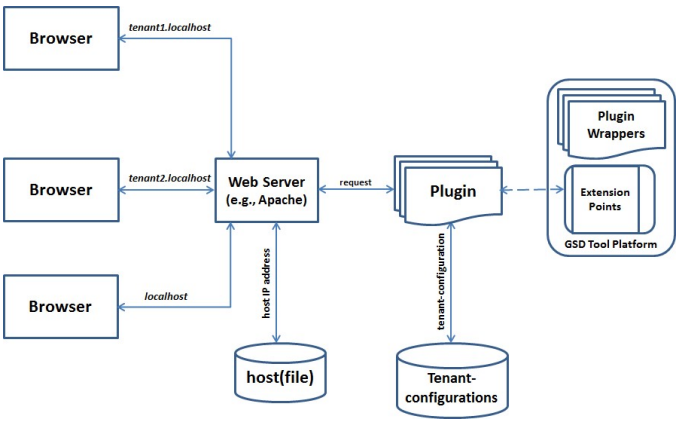


Fig. 1. COMITRE Architecture

We used a case study to evaluate the effect of tenant isolation at the data level during automated build verification/testing process for an application that logs every operation into a database in response to a specific event such as detecting changes in a file. To achieve this, we used Hudson's Files Found Trigger plugin, which polls one or more directories and starts a build if certain files are found in those directories [10]. Multitenancy isolation was implemented by modifying Hudson. This involved introducing a Java class into the plugin that accepts a filename as argument. During execution, the plugin is loaded into a separate class loader to avoid conflict with Hudson's core functionality [11].

To simulate multitenancy isolation at the data level when locking is enabled, we configured the data handling component in a way that isolates the data of different tenants (see Fig. 2). This is related to the concept of (i) locking is used in version control systems (e.g., Subversion) process to prevent clashes between multiple tenants operating on the same working copy of a file; and (ii) database isolation level which is used to control the degree of locking that occurs when multiple tenants or programs are attempting to access a database used by a cloud-hosted application. Most bugs/issue tracking applications (e.g., Bugzilla, ITracker, JIRA) use a database to store bugs [12]. Therefore, a tenant that first accesses an application component locks (or blocks) it from other tenants until the transaction commits.

3.2 Experimental Design and Statistical Analysis

A set of four tenants (T1, T2, T3, and T4) are configured into three groups to access an application component deployed using three different types of multitenancy patterns (i.e., shared component, tenant-isolated component, and dedicated component). Each pattern is regarded as a group in this experiment. We also created two different scenarios for all the tenants (see section 4.3 for details of the two scenarios). In addition, we also created a treatment for configuring T1 (see section 4.2 for details of the treatment). For each group, one of the four tenants (i.e., T1) is configured to experience a demanding deployment condition (e.g., large instant loads) while accessing the application component. Performance metrics (e.g., response times) and systems resource consumption (e.g., CPU) of each tenant are measured before the treatment (pre-test) and after the treatment (post-test) was introduced.

Based on this information, we adopt the Repeated Measures Design and Two-way Repeated Measures (within between) ANOVA for the experimental design and statistical analysis respectively. Experiments using repeated measures design make measurements using only one group of subjects, where tests on each subject are repeated more than once after different treatments [13]. The ***aim of the experiment*** is to evaluate the effect of locking on multitenancy isolation for components of cloud-hosted services. The ***hypothesis*** we are testing is that the performance and system's resource utilization experienced by tenants accessing an application component deployed using each multitenancy pattern changes significantly from the pre-test to the post test.

3.3 Experimental Setup and Procedure

The experimental setup consists of a private cloud setup using Ubuntu Enterprise Cloud (UEC). UEC is an open-source private cloud software that comes with Eucalyptus. The private cloud consists of six physical machines- one headnode and five sub-nodes. We used the typical minimal Eucalyptus configuration where all user-facing and back-end controlling components (Cloud Controller(CLC), Walrus Storage Controller, Cloud Controller (CC), and Storage Controller (SC)) are grouped on the first machine, and the Node Controller (NC) components are installed on the second physical machine. In our experiment, we installed NCs on all the other machines in order to achieve scalability for this configuration.

We use a remote client machine to access the GSD tool running on the instance via its public IP address. Apache JMeter is used as a load balancer as well as a load generator to generate workload (i.e., requests) to the instance and monitor responses. A file is pushed to a Hudson repository to trigger a build process that executes an Apache JMeter test plan configured for each tenant. Each instance is installed with SAR tool (from Red Hat *sysstat* package) and Linux du command to monitor and collect system activity information. Every tenant executes its own JMeter test plan which represents the different configurations of the multitenancy patterns.

To simulate multitenancy at the data level using JMeter, we use the JMeter Beanshell sampler to invoke a custom Java class that runs a query that sets the database transaction isolation level to SERIALIZABLE (i.e., the highest isolation level). To measure the effect of tenant isolation, we introduce a tenant that experiences a demanding deployment condition. We configured tenant 1 to simulate a *large instant load* by: (i) increasing the number of the requests using the thread count and loop count; (ii) increasing the size of the requests by attaching a large file to it; (iii) increasing the speed at which the requests are sent by reducing the ramp-up

period by onetenth, so that all the requests are sent ten times faster; and (iv) creating a heavy load burst by adding the Synchronous Timer to the Samplers in order to add delays between requests, such that a certain number of the request are fired at the same time. This treatment type is similar to unpredictable (i.e., sudden increase) workload and aggressive load.

Each tenant request is treated as a transaction composed of the 2 types of request: HTTP request and JDBC request. HTTP request triggers a build process while JDBC request logs data into the database which represents an application component that is being shared by the different tenants. Transaction controller was introduced to group all the samplers in order to get a total metrics (e.g., response) for carrying out the two requests. Figure 5 shows the experimental setup used to configure the test plan for the different tenants in Apache JMeter.

The initial setup values for experiment are as follows: (1) No of threads = 10 for tenant 1 (i.e., the tenant experiencing high load), and 5 for all other tenants; (2) Thread Loop count = 2; (3) Loop controller count = 10 for HTTP requests of tenant 1, and 5 for all other tenants; 200 for JDBC requests of tenant 1, and 100 for all other tenants; (4) Ramp-up period of 6 seconds for tenant 1 and 60 seconds for all other tenants; and (5) Estimated total number of expected requests = 250 for HTTP requests and 2500 for JDBC requests. This means that in each case the tenant experiencing high load receives two times the number of requests received by each of the other tenants. In addition, the requests are sent 10 times faster to simulate an aggressive load.

We performed 10 iterations for each run and used the values reported by JMeter and System activity report (SAR). The following system metrics were collected and analysed:

(i) CPU Usage: The %user values (i.e., the percentage of CPU time spent) reported by SAR were used to compute the CPU usage.
(ii) System load: We used the one-minute system load average reported by SAR.
(iii) Memory usage: We used the kbmemused (i.e., the amount of used memory in kilobytes) recorded by SAR.
(iv) Disk I/O: The disks input/output volume reported by SAR was recorded.
(v) Latency: The 90% latency reported by JMeter.
(vi) Throughput: We used the average throughput reported by JMeter.
(vii) Error %: The percentage of request with errors reported by JMeter.

4 Results

In this section, we discuss how the experimental results were analysed. We first performed A two-way (within-between) ANOVA to determine if the groups had significantly different changes from Pre-test to Post-test. Thereafter, we carried out *planned comparisons* involving the following: (i) a one-way ANOVA followed by Scheffe post hoc tests to determine which groups showed statistically significant changes relative to the other groups. The Dependent variable used in the one-way ANOVA test was determined by subtracting the Pre-test from Post-test values.

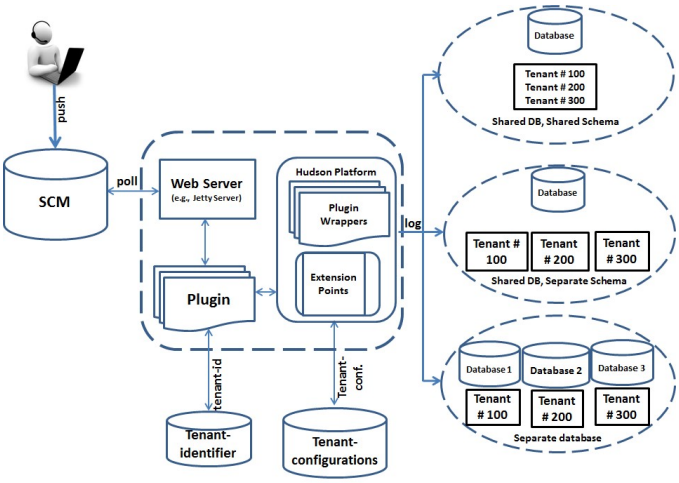


Fig. 2. Multitenancy Data Isolation Architecture

(ii) a paired sample test to determine if the subjects within any particular group changed significantly from pre-test to posttest measured at 95% confidence interval. This would give an indication as to whether or not the workload created by one tenant has affected the performance and resource utilization of other tenants. We used the "Select Cases" feature in SPSS to select the three tenants (i.e., the T2,T3,T4 that did not experience large instant loads) for each pattern.

Table 1 summarizes the effect of Tenant 1 (i.e., the tenant that experiences high load) on the other three tenants (T2, T2, T4). The key used in constructing the table is as follows: YES - represents a significant change in the metrics from pretest to post -test. NO - represents some level of change which cannot be regarded as significant; no significant influence on the tenants. The symbol "-" implies that the standard error of the difference is zero and hence no correlation and t-test statistics can be produced. This means that the difference between the pre-test and post-test values are nearly constant with no chance of variability. In the following, we present a brief discussion the findings of the study based on the estimate of the marginal means of change and paired sample test for scenario 1 and scenario 2.

(1) Response times and Error%: The paired sample test result shows that the response times of tenants changed significantly only for the dedicated pattern. A further analysis of the EMMC showed that the dedicated pattern had a much larger magnitude of change than all the other patterns. The Error% showed that there was no significant change in the tenants within any of the patterns; there was either no significant difference or no variability.

(2) Throughput: The results of the paired sample test showed that the tenants within all the patterns changed significantly from pre-

test to post-test. The shared component showed the smallest magnitude of change based on the plots of the EMMC.

(3) CPU: The plots of the EMMC showed that the shared component had the largest magnitude of change. The other two patterns were nearly the same. The paired sample test showed that shared component was the only pattern that changed significantly.

(4) Memory: The plot of the EMMC showed that the shared component changed showed the smallest magnitude of changed. We noticed an interesting trend in the sense that magnitude of change decreased steadily from the shared component to the dedicated component. The paired sample test showed that tenants deployed based on all the patterns changed significantly.

(5) Disk I/O: The paired sample test showed that there was no significant change between the tenants deployed based on the shared pattern. The plots of the Estimated Marginal Means of changed (EMMC) confirmed that the shared component changed the least.

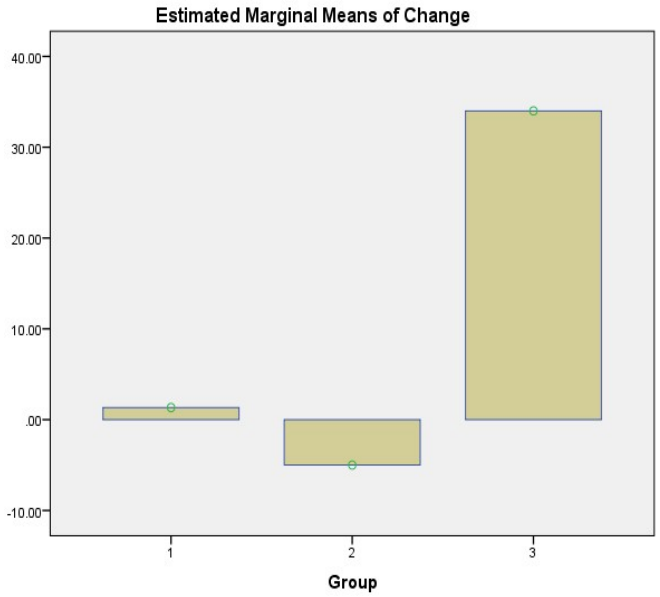


Fig. 3. Changes in response time

(6) System Load: The paired sample test showed that there was no significant influence on the system load for all the patterns. This means that even when locking is enabled the system load is not likely to change much.

5 Discussion

(1) CPU: The results showed that the CPU did not change significantly, except for the shared component. This implies that apart from the shared component, the degree of isolation was high. Therefore, we can say that although locking for enabled, there appears to be little or no influence in terms of resource consumption. This is understandable because Hudson, like many builders, do not consume much CPU.

(2) System Load: As the results show, the system load of the tenants showed either a nearly constant magnitude of change or no chance of variability. This means that even when locking is enabled, there may be no significant change in the system load as long as the size of the processor is large enough to cope of the number of piled-up requests.

(3) Memory: Builders are well known to consume a lot of memory, especially when handling difficult and complex builds. As the results showed, there was a significant difference between the tenants for all the patterns when locking was enabled. Overall, this means that there was a low degree of isolation between the tenants. In terms of the magnitude of change, the plots of EMMC showed the largest magnitude of change while dedicated component was the smallest. This implies that while the shared component is not recommended to minimize performance, but it may be used optimize the memory usage. On the other hand, the dedicated component can be used to avoid performance interference.

(4) Disk I/O: Compilers and builders generally consume a lot of disk I/O and it interacts directly with the operating system or the filesystem of the cloud platform used. As shown in the paired sample test result, tenants deployed based on shared component did not change significantly, implying a high degree of isolation. Therefore, when locking is enabled on an application component that is shared while carrying out I/O intensive builds, then the shared component would be recommended. The plots of the EMMC, confirms this position in the sense that the shared component showed the smallest magnitude of change out of the three patterns.

Table I. Paired Samples Test Analysis of Multitenancy Isolation When Locking is enabled

Pattern	Response times	Error%	Throughput	CPU	Memory	Disk I/O	System Load
Shared	No	No	Yes	Yes	Yes	No	-
Tenant-isolated	No	-	Yes	No	Yes	Yes	-
Dedicated	Yes	-	Yes	No	Yes	-	-

(5) Response times and Error%: The results show that the dedicated component had the largest magnitude of change for response times, while the reverse was the case for error% which had the largest magnitude of change for the shared component. This means that the shared component would not be recommended for preventing performance interference. It also shows that there would be a high possibility of requests timing out for tenants deployed based on shared component than for other tenants. A possible explanation for this is that requests can be delayed or blocked while trying to gain access to the shared application component.

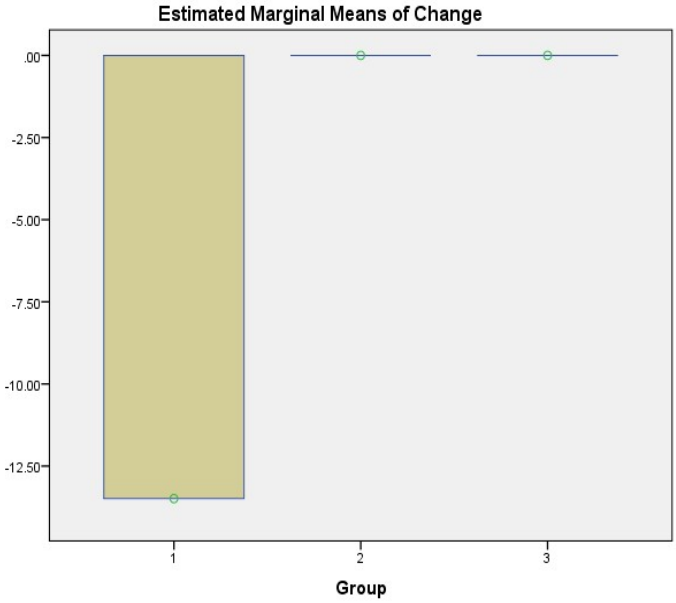


Fig. 4. Changes in error%

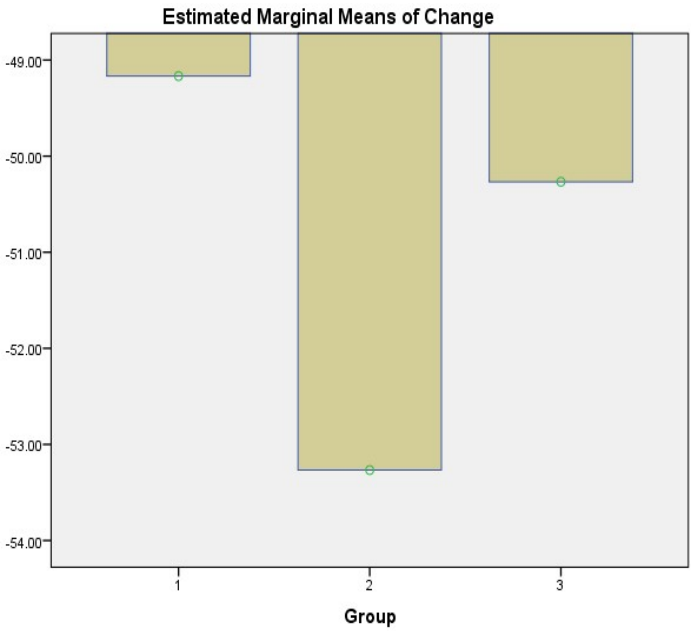


Fig. 5. Changes in throughput

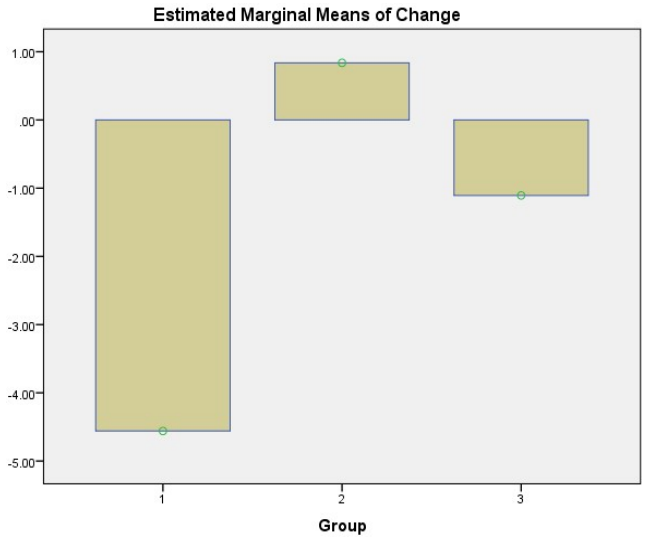


Fig. 6. Changes in CPU

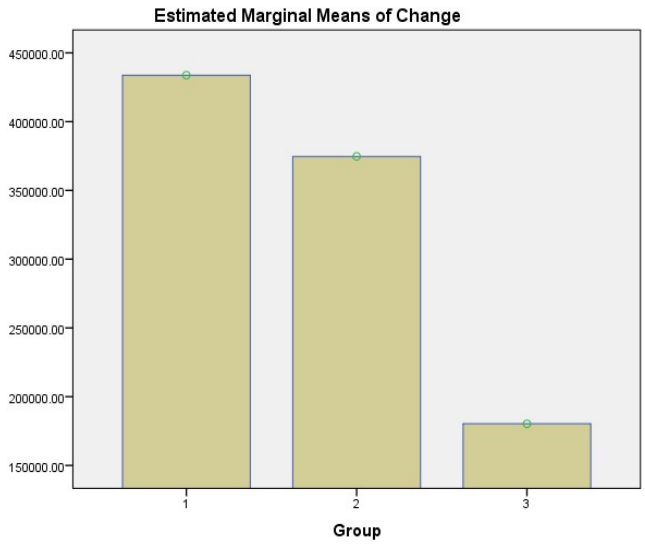


Fig. 7. Changes in memory

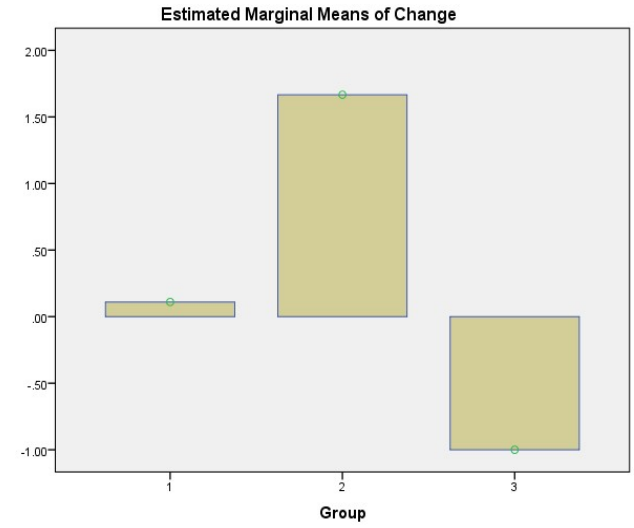


Fig. 8. Changes in disk I/O

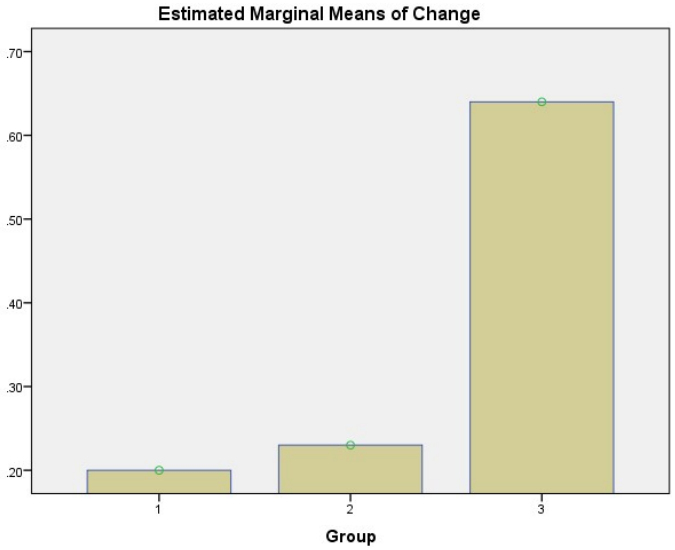


Fig. 9. Changes in system load

6. Recommendations and Limitations

The experimental results show that locking could have a significant effect on multitenancy isolation. Running a complete integration build in a slow network environment could take a lot of time and resources. To achieve the required degree of isolation, we recommend splitting the integration build into different stages and implement separate multitenancy patterns for each phase. For example, we could (i) creating a commit build that compiles and verifies the absence of critical errors when each developer commits changes to the main development stream based; and (ii) creating a secondary build(s) to run slower and less important tests. This study assumes that a small number of tenants send multiple requests to an application component deployed on a private cloud. The number of requests sent to the application component configured within Hudson was within the limit of the UEC private cloud used. Therefore, the results of this study should not be generalized to large public clouds.

7. Application of Locking on Cloud-hosted Software Development Tools and Associated Processes

A well-managed locking strategy is required to deal with real-time tightly synchronized/consistency-critical cloud applications such as such graph processing, financial applications, and real-time enterprise analysis applications. These cloud-hosted applications rely heavily on key software development processes such as continuous integration, version control and bug/issue tracking to build, test, and release software faster and more reliably.

Lock management in a multitenant cloud-hosted application is essential because if an architect misses placing a lock where required, then safety is violated. In contrast, if an architect inserts unneeded locks in a cloud-hosted application, then the performance of the system suffers due to the unnecessary synchronizations [14]. In the following, we discuss how locking is used in three important types of software development processes, and some recommendations to follow regarding achieving the required degree of multitenancy isolation.

7.1. Locking in Continuous Integration process

Locking is a very important operation in a typical continuous integration process. For example, in Hudson, it is used to block builds dependencies from starting if an upstream or downstream project is in the build queue. One implication of this is that if there is a presence of piled-up requests/builds on the queue, then the system load is likely to be affected. This was not the case in the experiments and so the system load was nearly constant with no chance of variability.

We recommend that in order to optimize resources that support a cloud-hosted service while at the same time guaranteeing multitenancy isolation, the architect should avoid certain operations lock processes for a long time, especially when there is either limited resources or frequent workload changes. Such operations include carrying out difficult and complex builds (i.e., builds that have many interdependencies with other programs or systems), and (ii) running a large number of builds concurrently.

7.2. Locking in Version Control process

Locking (similar to the “reserved checkouts” mechanism) is used internally in version control process (e.g., in Subversion) to achieve mutual exclusion between users to avoid clashing commits or to prevent clashes between multiple tenants operating on the same working copy. A Version control system can be setup to use a database as its backend. For example, it is common for architects to setup subversion to store data in a Berkeley DB database environment. When this is the case, locking can be used internally by the Berkeley DB to prevent classes between multiple processes and programs trying to access the database.

With respect to multitenancy isolation, when multiple tenants are accessing a shared version control repository, it implies a shared component is being used for deployment. Under this situation, it is possible for fatal errors or interruptions to occur which can prevent a process from having the chance to remove the locks it has placed in the database. While implementing dedicated component deployment would be an obvious solution to avoid such interferences, one would have to go a step further when working with networked repository. This could involve putting in place an off-site backup strategy, and shutting down server programs (e.g., Apache HTTP server) from accessing or attempting to access the repository.

When using a version control system such as subversion that implements locking, fetching large data remotely and finalizing a commit operation can lead to unacceptably slow response times and can even cause tenants request to time out. Therefore, having the repository together with the working copy located on your machine is beneficial. It is also important to note that file locking

along with data compression are some of the operations that could consume resources, especially when accessing a shared repository from a client with a slow network and low bandwidth.

7.3. Locking in Bug tracking process

A bug tracking system is used to keep track of reported software bugs in software development projects. A major component of a bug tracking system is the storage component that records facts about known bugs. Depending on the type of storage component used to store bugs, locking can be used to prevent multiple tenants trying to access the bug data store.

Most bug and issue tracking systems (e.g., Bugzilla and JIRA) use a database to store bugs. Enabling locking on the bug database, for example, can also increase resource consumption (e.g., CPU, memory), especially when running long transactions, running complex transactions concurrently or transferring large bug attachments across a slow network connection.

8. Conclusion and Future Work

In this paper, we have presented the effect of locking on multitenancy isolation for components of a cloud-hosted service to contribute to literature on multitenancy isolation and cloud deployment of application components. The study revealed that when locking is enabled for components of a cloud-hosted service, it can have a significant impact on the performance and resource consumption of tenants especially for operations that interact directly with the local file system (e.g., FAT, NTFS, GoogleFS**,** HFS+) of the platform on which the service is hosted. One option we have recommended is to split a software process (e.g., a long build process) into separate phases and then implement different degrees of isolation for each phase.

We plan to apply our approach to implementing multitenancy isolation for a cloud-hosted service in a distributed scenario where locking is enabled for all or some of the components at different of the cloud stack. For example, in distributed bug tracking some bug trackers like Fossil and Veracity are either designed to use (or integrated with) distributed VC or CI systems, thus allowing bugs to be created automatically and inserted to the database at varying frequencies.

References

[1] C. Fehling, F. Leymann, R. Retter, W. Schupeck, and P. Arbitter, *Cloud Computing Patterns*. Springer, 2014.

[2] E. Bauer and R. Adams, *Reliability and availability of cloud computing*. John Wiley & Sons, 2012.

[3] B. Collins-Sussman, B. Fitzpatrick, and M. Pilato, *Version control with subversion*. O'Reilly, 2004.

[4] M. Moser and T. O'Brien. The Hudson book. Oracle, Inc., USA. Online: accessed in November, 2017 from http://www.eclipse.org/hudson/the-hudsonbook/book-hudson.pdf.

[5] Bugzilla.org. The bugzilla guide. The Mozilla Foundation. [Online: accessed in November, 2017 from http://www.bugzilla.org/docs/.

[6] R. Krebs, C. Momm, and S. Kounev, "Architectural concerns in multi-tenant saas applications." *CLOSER*, vol. 12, pp. 426–431, 2012.

[7] S. Strauch, V. Andrikopoulos, F. Leymann, D. Muhler, "Esbmt: Enabling multi-tenancy in enterprise service buses," *CloudCom*, vol. 12, pp. 456–463, 2012.

[8] S. Walraven, T. Monheim, E. Truyen, and W. Joosen, "Towards performance isolation in multi-tenant saas applications," in *Proceedings of the 7th Workshop on Middleware for Next Generation Internet Computing*. ACM, 2012, p. 6.

[9] L. C. Ochei, J. Bass, and A. Petrovski, "Evaluating degrees of multitenancy isolation: A case study of cloud-hosted gsd tools," in *2015 International Conference on Cloud and Autonomic Computing (ICCAC)*. IEEE, 2015, pp. 101–112.

[10] Hudson. Apache software foundation. [Online: accessed in January 2017 from http://wiki.hudsonci.org//display/HUDSON/Files+Found+Trigger].

[11] L. C. Ochei, A. Petrovski, and J. Bass, "Evaluating degrees of isolation between tenants enabled by multitenancy patterns for cloud-hosted version control systems (vcs)," *International Journal of Intelligent Computing Research*, vol. 6, Issue 3, pp. 601 – 612, 2015.

[12] Serrano, N. and Ciordia, I., 2005. Bugzilla, ITracker, and other bug trackers. *IEEE software*, *22*(2), pp.11-13.

[13] Verma, J.P., 2015. Repeated measures design for empirical researchers. John Wiley & Sons.

[14] Demirbas, M., Tasci, S. and Kulkarni, S., 2012, July. Maestro: A cloud computing framework with automated locking. In *Computers and Communications (ISCC), 2012 IEEE Symposium on* (pp. 000833-000838). IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 97-104 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Novel Technique for Enhancing Color of Undersea Deblurred Imagery

Chrispin Jiji*, Nagaraj Ramrao

Department of Electronics and Communication, The Oxford College of Engineering, Oxford Institutions, Bangalore, India

ARTICLE INFO

Article history:
Received: 18 September, 2018
Accepted: 23 October, 2018
Online: 01 November, 2018

Keywords:
Underwater
Image Deblurring
Image Enhancement

ABSTRACT

Exploring the ocean underneath has always been an area of great scientific and environmental concern. However, the study of underwater environment was very difficult due to the extreme conditions. Undersea descriptions undergo severe distortion attributed to absorptive as well as scattering properties. Absorption substantially removes illumination, whereas a ray of light redirected in several path when it interacts by substance. Because of these, undersea descriptions encompass blur as well as color loss. In this paper we suggested an effective technique namely, a turbidity removal method for deblurring the image. If the deblurred image has a lighting problem, we make use of a color-correction method to find the clear image. Our substantial qualitative and quantitative assessment expose that the proposed algorithm progress the excellence as well as lessen color distortion loyally, also improves the state-of-the-art undersea technique.

1. Introduction

Images captured in undersea has plays a vital basis of interest within various branches of technical and systematic explores [1], such as examining underneath infrastructures [2] as well as cables [3], detecting manmade objects [4], managing undersea vehicle [5], marine biology investigate [6], and archaeology [7]. Apart of normal descriptions, undersea descriptions undergo reduced visibility ensuing attenuation of the propagated illumination, mainly owing to absorption along with scattering effect. In this paper, we use image processing acting extensive interest over earlier years due to its challenging nature and its importance for the surroundings. Improving undersea scene excellence separates the problem into image restoration and image Enhancement

Visibility in undersea imagery is usually blurry, but having large number of particles underneath cause's cloudiness or more haziness, called turbidity, which causes blur in undersea imagery. To remove any blur, we usually use restoration problem with estimated or known PSF matrix.

$$b = h * d + n \quad (1)$$

where d represents deblurred representation, h denotes PSF kernel, and b denotes turbid representation. The main challenge is estimating blur kernel [8-12]. The blur kernel cannot be estimated directly as it varies depending on blur itself. Existing system uses some prior to estimate point spread function (PSF) for restoration of undersea blurred imagery.

If the deblurred image has any difficulty in lighting, we use image enhancement to enhance the picture excellence. The difficulty during lighting owing towards absorption substantially reduces illumination, practically, in undersea images shown in Figure 1. The objects by remoteness more than 10m are about imperceptible, and colors go down by the deepness of water.

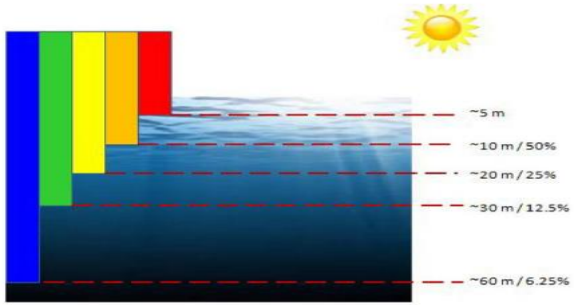


Figure 1: Dissimilar illumination weakened by dissimilar charge

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Do not use abbreviations in the title or heads unless they are unavoidable. In general, red illuminations vanish by deepness of 5m, after that brown followed by yellow radiance, lastly green as well as blue illumination vanish by deepness with 30m as well as 60m. Thus undersea imagery is subject to blue-green color which changes the picture excellence. As a result, we used color adjustment scheme to compensate light condition. There have been several attempts for restoration as well as enhancement, since degraded picture

*A. Chrispin Jiji, 8951627124 & chrispinjij@gmail.com

outcomes the understanding of multiplicative by means of additive process [13]. Conventional enhancement methods namely gamma correction, histogram equalization strongly restricted for such a task. This difficulty deal with modified attainment approach via various imagery [14], specific hardware [15] or polarization filters [16]. Even though their important attainment, these approach undergo numeral concern that decrease their convenient applicability.

As shown in Figure 2, our technique uses two stages namely image restoration stage for deblurring and enhancement stage for enhancing its excellence. In this paper we suggest an adaptive sparse domain selection (ASDS) system to restore the image. By training compact sub-dictionaries from high quality example image cluster the window. Since each cluster employ several windows by related prototype, compact sub-dictionary learns for every cluster. Particularly, for simplicity we use principal component analysis (PCA) technique towards learning sub-dictionaries. The most excellent sub-dictionary that is mainly applicable to given window is chosen, because the given window is better represent sub-dictionary is accurately reconstruct entire image. Besides sparsity regularization, other terms as well initiate for improving its performance. Later use autoregressive (AR) models, pre-learned from training dataset characterizing confined structures. For every confined window, we choose the AR model. On the other hand, considering fact that there are often several repetitive image structures in an image, we introduce a non-local (NL) self-similarity constraint served as another term, helpful for preserving boundary sharpness and restrain noise. After introducing ASDS using two constraints (AR & NL) into IR structure, we present a weighted Gray Edge method for solving lighting problem. Substantial experimentation on image deblurring and enhancement show that the projected approach effectively reconstructs picture details along with excellence, outperforming different state-of-the-art IR with IE methods in terms of both excellence metrics in addition to ocular insight.

The paper is planned as follows. Section 2 provides concise outline about earlier art. Section 3 signifies our restoration approach, about turbidity removal method especially used for undersea descriptions. Section 4 describes novel enhancement approach, mainly to improve its excellence. Section 5 presents comparative qualitative and quantitative estimation of undersea system and Section 6 provides closing comments.

2. Previous Art

This part reviews main advances to deblur or else improve imagery confined undersea. In computer vision, methods to handle ambiguity is roughly on some knowledge or assumptions known beforehand, *i.e.*, *priors*. The priors impose extra constraints/dependency among the unknown variables. In the following, we review the previous turbidity removal methods. We do not discuss the technical details among those methods. Instead, we concern about the extra constraints. All methods reformulate in a same framework expressed indifferent forms in the original works. Single image turbidity removal methods have to rely on some priors.

The prior is statistical/physical properties, heuristic assumptions, simplifications, along with application-based rules. The blur imaging model in (1), discrepancy between some equations and unknown, the prior expected to introduce at least one constraint for each pixel. The challenge of recovering d from b is under-constrained. To make it solvable, extra knowledge has been built-into the restoration method. The former understanding is often built-in with a regularization term, principal to the later energy minimization problem:

$$\hat{d} = \arg\min\left\{\left\|b - Hd\right\|_2^2 + \lambda J\left(d\right)\right\} \qquad (2)$$

where λ denotes Lagrangian multiplier matching the exchange among former term $J\left(d\right)$ along with likelihood $\left\|b - Hd\right\|_2^2$. The former term $J\left(d\right)$ act as a main part in the restoration method. Sparsity based IR process guess that the natural representation is sparse in few fields. Sparse representation lately has paying attention to investigators for resolving complications including deblurring, denoising, super resolution. In [17], the ARM prototypes are in the neighborhood calculated commencing an originally expected image besides bring about much better-quality for TV prior trendy reforming boundary associations. Now, we resolve to suggest knowledge established using adaptive prior, where the AR prototypes remain knowledgeable commencing great feature of training imageries, towards raising AR forming accuracy. The approach presents exact modest: the blotches that make sure alike forms can exist spatially distant, besides gather the entire appearance. In [18-20], the NLM prior combined through the sparse domain, as of alike appearance blotches be concurrently implied towards the strength of converse reestablishment.

Image enhancement based methods are not required to solve the physical form, but rather directly enhance its contrast as well as improving its excellence from human visual perception. Within undersea, color is extremely related by deepness, and a significant difficulty is green-bluish form desires to resolve. Since as the illumination go through undersea, reduction process affects wavelength spectrum, thus affecting the gray level along with appearance of colored surface. Existing white balancing process is a key to our domain. Next, we briefly change those techniques and give details for our novel approach projected by undersea imagery. Generally a scheme formulates exact guess to estimate color, in addition to meet color constancy via color channel will standardize the radiance. Projected scheme use a statement that majority of patch show off extremely small gray level in at-least single color channel that openly about blur mass and progress vivid colors. It cannot sufficiently deal with color deformation and complex arrangement. The deblurred imagery will attribute color swing with artifact effect. The Grey World system [21] believe that the averages of the three color constituent, illuminate during impartial radiance basis further acquire poorer color patch. The Max-RGB [21] resting on the RGB color. Shades-of-Grey [22] use Minkowski norm-p is achromatic. The Grey Edge method [23] employs items boundary data towards adjusting color fidelity; it assumes that normal derivative of color mechanism is achromatic. Every color provides pre procedure by Gaussian filter is typically 1-2 toward determining the statement. It might worsen the performance. Weighted Grey Edge method [24] use boundary data of variety of substance, intended for instance, shadow and mechanism. This paper is an extended work initially presented during ICPCSI [25]. Also the journal paper intended novel technique for enhancing undersea imagery. Revised result much

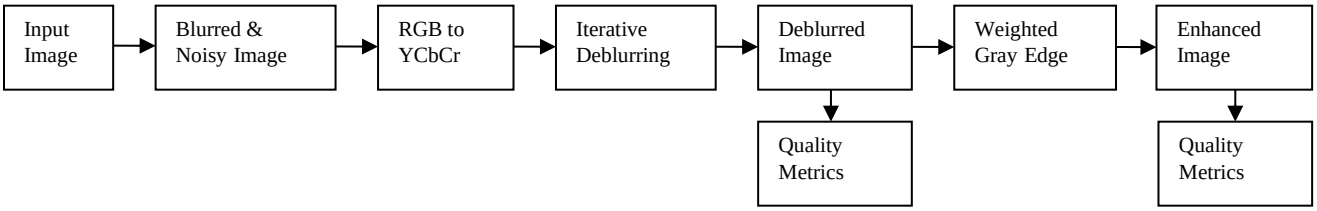


Figure 2: Block diagram of Proposed Method

improves the lighting problem in undersea imagery. Enhancement based methods not required to solve physical form of degraded image, but rather directly enhance contrast and improve image quality from human visual perception.

3. Underwater image restoration for turbidity removal method

Underwater turbidity removal is very challenging with essential part of picture suffers from turbidity. Several algorithm deals with turbidity elimination process recognized as Image deconvolution (ID). To get back the undersea deblurred picture, we used [26-27]. Here we adaptively study compacted sub-dictionaries to every confined patch. Let $d_m \in R^n$ denote the image block dimensions √n x √n attained from deblurred picture, we can describe $d_m = R_m D, m = 1,2,.....N$ wherever R_m denotes matrix from d_m. Rendering towards sparse prior d_m can be signified by means of outmoded dictionary, namely $d_m = \phi_{k_m} \alpha_m$, thus D can be conveyed by way of

$$D = \left(\sum_{m=1}^{N} R_m^T R_m \right)^{-1} \sum_{m=1}^{N} R_m^T \phi \alpha_m \tag{3}$$

For convenience, we define:

$$D = \phi o \alpha = \left\{ \sum_{m=1}^{N} R_m^T R_m \right\}^{-1} \sum_{m=1}^{N} R_m^T \phi \alpha_m \tag{4}$$

Now, the primary restored image denoted as $D^{(k)}$ constantly drop some information in the input image and comprises some artifacts. In order to further appropriate for human visual scheme, we update the restored image in every iteration.

For each patch we adaptively choose the sub-dictionary based on least distance specified through

$$k_m = \left\| d_m^h - \mu_k \right\|_2^2 \tag{5}$$

3.1 Adaptively Reweighted Sparsity Regularization

Existing work uses λ=1 for whole patches. In our work λ differs from patch to patch to expand the excellence of restored image. The reweighted sparsity regularization is specified in place of

$$\hat{\alpha} = \left\{ \left\| b - H \phi o \alpha \right\|_2^2 + \sum_{m=1}^{N} \sum_{n=1}^{o} \lambda_{m,n} \left| \alpha_{m,n} \right| \right\} \tag{6}$$

Then weight λ_{mn} stands as

$$\lambda_{m,n} = \frac{2\sqrt{2}\sigma_{n1}^2}{\sigma_{m,n} + \varepsilon} \tag{7}$$

Reweighted procedure is as follows:

Step 1	Fix ℓ = 0 as well as $\lambda_n^l = 1$
Step 2	Resolve the subjective l_1 minimization problematic $d^{(l)} = \arg\min \left\| \lambda_n^l \phi^T d \right\|_1$ Focused towards $d = \phi \alpha^{(l)}$
Step 3	Set $\alpha^{(l)} = \phi^T d^{(l)}$ and describe $\alpha_n^{(l+1)} = \frac{2\sqrt{2}\sigma_{n1}^2}{\sigma_{m,n} + \varepsilon}$
Step 4	Proceed until l reaches $l_{\max}$ Else, increase l and drive towards step 2

3.2 Auto Regressive model

The ASDS progress extensively by the usefulness of sparse modeling and later outcome image deblurring. To further get better excellence of deblurred imagery, we established autoregressive (AR) models to formalize the image confined smoothness. For each high pass filtered patch adaptively choice AR dictionary by means of calculating the Euclidean distance. For each present pel we make sure four neighbouring pels and its weights are specified as;

$$w_{m,n} = \frac{1 / \left(\left\| d_m^h - \mu \right\|_2^2 + \varepsilon \right)}{\sum 1 / \left(\left\| d_m^h - \mu \right\|_2^2 + \varepsilon \right)} \tag{8}$$

Therefore, the weighted sum of entire adjacent pels expressed as

$$\chi_m = \sum w_{mn} a_{k_m} \tag{9}$$

Assemble the exceeding matrix into 3x3 window we resolve the value of local geometry model. Let d_m remains the center pel, and χ_m stays the vector comprising of the adjacent pels which is adjacent to middle pel d_m, now the best middle pel value would lessen $\left\| d_m - a_{k_m}^T \chi_m \right\|_2^2$ .For the ease of expression we write $\sum_{d_i \in d} \left\| d_m - a_{k_m}^T \chi_m \right\|_2^2$ as $\left\| (I - A) \phi o \alpha \right\|_2^2$. By integrating this restraint, the restoration equation as follows

$$\hat{\alpha} = \left\{ \left\| b - H\phi o\alpha \right\|_2^2 + \sum_{m=1}^{N} \sum_{n=1}^{o} \lambda_{m,n} \left| \alpha_{m,n} \right| + \gamma \left\| (I - A)\phi o\alpha \right\|_2^2 \right\} \quad (10)$$

where γ represents a constant balance the role of AR formalization term.

3.1 Nonlocal Self Similarity model

The AR model utilizes the confined information in every patch. On the other hand, there are often numerous cyclical prototypes all through the representation. Thus non-local redundancy mainly improves the excellence of imagery. Along with AR models an additional term called non-local similarity added into the IR structure. The source of global mean filtering is very simple, aimed at current image block; we catch numerous alike blocks to restrain it. Aimed at each image block d_m, we discover all image blocks alike to it in the complete image.Take on d_m^s is one of blocks similar to d_m, then $e_m^s = \left\| d_m^s - d_m \right\|_2^2$ would be small enough. If we estimate the summation of entire midpoint pels voguish these alike blocks, besides effects would satisfy $d_m \approx \sum_{s=1}^{L} b_m^s x_m^s$ where b_m^s denotes the weight owed to d_m^s.

Apparently, the more two blocks are alike; the bigger weight would be assigned. Thus, we can define the expression for weight calculation as $b_m^s = \exp\left(-e_m^s / h\right) / \sum \exp\left(-e_m^s / h\right)$ where h stays a constant directing the window shape. The global model in an image can be estimated such as

$$\sum_{m=1}^{N} \left\| d_m - \sum_{s=1}^{L} b_m^s d_m^s \right\|_2^2 \quad (11)$$

Mathematically in form of $\left\| (I - B)\phi\alpha \right\|_2^2$. Besides incorporate equally the AR regularization and the NLSS based sparse depiction in Eq. (12), thus solves the problem using AR to regularize image confined smoothness with NLSS towards utilizing the image Non-Local redundancies as given as

$$\hat{\alpha} = \left\{ \left\| b - H\phi o\alpha \right\|_2^2 + \sum_{m=1}^{N} \sum_{n=1}^{o} \lambda_{m,n} \left| \alpha_{m,n} \right| + \gamma \left\| (I - A1)\phi o\alpha \right\|_2^2 + \eta \left\| (I - A2)\phi o\alpha \right\|_2^2 \right\} \quad (12)$$

Meanwhile η panels the stability among local adaptation plus nonlocal strength.

The projected technique advances the existing schemes into excellence metrics despite undersea picture is crucial to recover color. In next section, we therefore propose color adjustment approach, relying on weighted grey model. As depicted within Figure 2, our approach adopts a novel strategy, for compensating color cast so as to enhance detail scene.

4. Underwater image enhancement of deblurred image

The deblurred image is subject to global enhancement with no prior required. This is necessary to lessen the architects due to deblurring, and for enhancement of color information otherwise degraded in the underwater image formation process. Different depth levels allow different colors to distort. Thus the faded as well as non-uniform color distribution will describe the undersea representation. Mauricio [28, 29] proposed an algorithm to readily pre-process the underwater imaging using homomorphic filtering, wavelet de-noising anisotropic filtering, and RGB color channel equalization to enhance color. The technique is usual and need no prior or constraint change. Another such scheme uses distribution of the histogram to make the bins change according to color density. In the proposed work, we use a weighted grey edge scheme to improve the color constancy.

4.1 Color adjustment section

Color adjustment is crucial in undersea, we apply our system towards deblurred imagery thus improving the picture look by means of neglecting redundant color casts caused by diverse radiance. To obtain our output we achieve a gamma correction method which endeavours to correct global dissimilarity; hence progress color constancy via weighted gray edge. It dispenses dissimilar mass to diverse boundaries according to the information of boundaries called weighted grey edge [24]:

$$\left(\int \left| w(f_t)^K \; f_{t,D}(D) \right|^p dD \right)^{\frac{1}{p}} = Ki_f \quad (13)$$

Where $w(f_t)$ denotes mass purpose, K remains constant to impose the mass of the boundaries, $f(D)$ stays on input, i_f denotes final-illuminant to make proper color. This modifies colors as close as original colors, considerably improving ending outcome. Numerous scheme incorporates towards accurately estimating color, an appropriate weighting system must impose appropriate data about color and ignore unrelated data. Given that cue intended for estimating color, an obvious choice towards computing weight using specular edge detection methods.

4.2 Different boundary methods

A variety of boundary type as material, shadow or shading, specular and inter-reflection boundaries. i) Material boundaries are conversion among two dissimilar surface or else items. ii) Shading boundaries are conversion that reasons the geometry of an item, intended for an alteration in surface direction by means of light. It is also cause obstruct in the beam source. Obstruct intensity gradient, however occasionally introduce faint color gradient. iii) Inter-reflection is the cause of radiance show as of one surface against a second surface. Thus, further transform of entire lighting towards inmost near next surface, therefore color, perceived via subtracting the derivative of invariant as well as true image. The projected scheme includes various photometric descriptions. Further exclusively, quasi invariants use many flexible weighting systems, resultant within a well-designed weighting system. Quasi invariants calculate the derivative of a representation, with three photometric variants

$$f_D = \left(f_{R,D}, f_{G,D}, f_{B,D} \right)^T \quad (14)$$

Besides eliminating inconsistency, a set of derivatives making called quasi invariants. Using quasi invariants portray, three dissimilar weighting schemes.

4.2.1 Specular boundary system

The quasi invariants decay picture into three paths. Ridge of derived term resting on luminosity basis called specular variant which definite as

$$O_D = \left(f_D . \hat{C}^{\wedge i} \right) \hat{C}^{\wedge i} \tag{15}$$

where $\hat{C}^{\wedge i} = \frac{1}{\sqrt{3}}(1,1,1)^T$ denotes color source and dot signify vector product. The derivative reasons with places of interest and subtracted with variation as of its derivative termed as

$$O_D^t = f_D - o_D \tag{16}$$

It merely enclosed shadow shade with material boundaries and is not sensitive towards emphasizing boundaries. While the entire derivative power enclosed in three various paths, relation among specular variation vs whole sum of energy be a suggestion that a boundary is specular or else not. The specular weighting system as:

$$W_S(f_D) = \frac{|o_D|}{\|f_D\|} \tag{17}$$

Where $|o_D|$ describes complete rate with o_D in addition to $\|f_D\| = \sqrt{f_{R,D}^2 + f_{G,D}^2 + f_{B,D}^2}$.

4.2.2 Shadow boundary system

By means of similar analysis happening, the shadow-shading route an invariant in addition to quasi invariant attained as:

$$\begin{aligned} S_D &= \left(f_D . \hat{e} \right) \hat{e} \\ S_D^t &= f_D - S_D \end{aligned} \tag{18}$$

Where $\hat{e} = \frac{1}{\sqrt{R^2 + G^2 + B^2}}(R,G,B)^T$. This quasi-invariant is not sensitive towards shadow boundaries. Translating this yields the next effect:

$$W_{SD}(e) = \frac{|S_D|}{\|f_D\|} \tag{19}$$

4.2.3 Material boundary system

At last, shadow system along with quasi invariant creates extrapolate derivative taking place hue path:

$$\begin{aligned} H_D &= \left(f_D . \hat{b} \right) \hat{b} \\ H_D^t &= f_D - H_D \end{aligned} \tag{20}$$

where denotes hue path, perpendicular on the way to preceding paths:

$$\hat{b} = \frac{\hat{e} \times \hat{C}^{\wedge i}}{\left| \hat{e} \times \hat{C}^{\wedge i} \right|} \tag{21}$$

where H_D^t do not enfold specular or else shadow boundaries, which then employ better weights to objects boundaries resembling

$$W_M(f_D) = \frac{|H_D^t|}{\|f_D\|} \tag{22}$$

To assess the control of boundary -kind classifier on the color constancy outcome. Out of these three methods specular edge type used to estimate final illuminant.

4.3 Iterative Weighted Gray Edge

The projected scheme initially correct deblurred imagery by an approximated illuminant. Then, last color corrected output image is given as

$$f_f = f_D \left(\frac{1}{i_f} \right)^D \tag{23}$$

Subsequently, in suggested method, a novel weighting scheme is to compute color corrected imagery for every iteration. For clearness, we will not change the weighting scheme throughout the iterations. Additionally, the early light approximate white basis $(1,1,1)^T$ otherwise end the color adjusted result. At last, convergence defined for predestined several iterations.

5. Experiment Results

Experiments results focussed towards removing blur and enhancing color. The improved underneath picture illustrates rigorous color alteration; however not succeed to improve colors completely. Deblurring method in Figure.3 gives deblurred picture; furthermore Figure 4 yield enhanced result to improve the color. Therefore, the projected scheme yields improved performance than the conventional system via four major branches: (A) Underwater Deblurring Evaluation (B) Underwater Enhancement Evaluation, and (C) Quantitative Evaluation. These parts depict individual performances of each module.

5.1 Underwater Deblurring Evaluation

It is an effective scheme for reconstructing original representation. Estimated PSF is added with original to make blurry which then added with Gaussian noise to get a blurry representation, which subjected through iterative deblurring schemes to recover novel imagery. Our algorithm in Figure 3 applied to only luminance element because human vision is more responsive to luminance variation.

5.2 Underwater Enhancement Evaluation

The undersea turbidity removal imagery improves the eminence although it produces severe color deformation owing towards absorbing band with underneath element. Thus inaccurate statement is based on every hue band is evenly absorbed in underneath. Thus projected technique in Figure 4 employ weighted gray edge process towards make even every RGB means and thus resolve distortion problem captured in underneath. Assessment among preceding technique exposed in Figure 5. While our method can proficiently get away blur with color deformation to get proper color exclusive by artifacts and gamma adjustment to correct global contrast moreover undersea imagery tend to emerge too light. The connection among true, deblurred as well as color output illustrated in Figure 6.

5.3 Quantitative Evaluation

Color adjustment using final illuminant gives final output, used by computer vision systems in many applications. Then projected algorithms give improved results as evaluated to existing scheme [30]. Table summarize metrics using Image quality Assessment and unique Image quality Assessment (SIQA). Thus projected scheme be superior than conventional scheme gives improved result.

a) Ordinary IQA

It mainly assesses contrast and structural adjustment. The most commonly used IQAs introduced as follows:*Mean:* Mean is the sum of all values in matrix.

$$\mu = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x_{m,n} \tag{24}$$

b) Standard deviation (SD): It reflects the degree of dispersion within picture on its standard significance, and contrast in certain range. The larger SD, better the visual effect will be:

$$SD = \sum_{m=1}^{M}\sum_{n=1}^{N}\sqrt{\frac{\left(f(m,n)-\mu\right)^2}{MN}} \tag{25}$$

where *M* with *N* denotes row, column of imagery; $f(i,j)$ denoted intensity of pel moreover μ belongs to standard rate of whole image.

c) *Entropy (E):* An image taken as a source of random output sets $\{a_i\}$ and the probability with a_i is $P(a_i)$ then standard amount by data in image as shown as

$$H = -\sum_{i=1}^{L} P(a_i)\log_2 P(a_i) \tag{26}$$

Higher value of E, more the information in picture.

d) Mean squared error (MSE): The full-reference excellence metric computed via average squared intensity differences of distortion along with reference representation pels as

$$MSE = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left[f(m,n) - f^{'}(m,n)\right] \tag{27}$$

where *M* and *N* denotes row and column of the imagery, $f(m,n)$ remains original with $f^{'}(m,n)$ denotes deblurred picture.

e) Peak SNR (PSNR): It is used as index for signal distortion. Larger the PSNR smaller the distortion and expressed as:

$$PSNR = 10\log\frac{f_{\max}^2}{MSE} \tag{28}$$

where $f_{\max}$ stays on largest gray value, in general $f_{\max} = 255$.

f) *Structural similarity (SSIM):* Generally, the ocular view is extremely modified to extract data on picture, thus measures restored picture excellence using three mechanism specifically; luminance $L(D,E)$, contrast $c(D,E)$ along with structure comparison *S(D,E)*. All these combined to give up whole similarity computation as:

$$S(m,n) = F\left(L(D,E), c(D,E), S(D,E)\right) \tag{29}$$

The similarity of the two images is dependent on SSIM, and has a value between [0,1]. When the value is close to 1, the two images are more similar. SD reflects dissimilarity of the image; E reflects the data; later MSE, PSNR and SSIM reflects the degree of distortion. Higher MSE, lower PSNR and SSIM scores imply greater dissimilarity among enhanced results and referenced deblurred image. This measures often used for simple calculation enclose clear and suitable for optimization. Conversely, these approaches cannot be simply adopted, because preceding IQA metrics are usually unsuitable for relevance intend to measure distortion intensity rather than visibility in imagery.

5.3.2 Special IQA

Some IQAs designed particularly for image from different views as follows.

a) Image Visibility Measurement (IVM)

Inspired with blind assessment indicator, an extra picture visibility measurement employ perceptible boundary segmentation as

$$IVM = \frac{n_v}{n_{total}}\log\sum_{D\in\Im} C(D) \tag{30}$$

where n_v represents the amount of perceptible boundaries, n_{total} denotes amount of boundaries, $C(D)$ remains average dissimilarity, along with $\Im$ denotes picture region with perceptible boundary.

b) Histogram correlation coefficient (HCC)

A good deblurring technique should allow improved picture towards Histogram allocation. It employs dual color imagery as standard to measure the act of color enhancement.

c) Contrast gain

The dissimilarity of clear picture is much higher than that of deblurred imagery, that check dissimilar enhancing system. Higher contrast, better the scheme to be. Global dissimilarity is for evaluating different techniques. It signifies mean contrast comparison among enhanced as well as deblurred image by

$$C_{gain} = \bar{C}_J - \bar{C}_I \tag{31}$$

where $\bar{C}_J$ in addition to $\bar{C}_I$ represents mean contrast of enhanced and deblurred representation, respectively.

d) Ocular dissimilarity assess (ODS)

It compute the amount of perception using

$$ODS = 100 * R_v / R_t \tag{32}$$

where R_v represents amount with limited region, SD is higher than known boundary with R_t remains entire amount of limited region. We chose OTSU image segmentation algorithm that adaptively compute the limit. It uses local standard deviation that denotes dissimilarity picture towards determining visibility.



(i) (ii) (iii) (iv)

Figure 3: Underwater Turbidity Removal. (i)Blurred picture (ii) ASDS (iii) ASDS using AR (iv) ASDS using AR and NL



a) Deblurred b) grad_im c) weight_map d) Ours

Figure 4: Proposed Method



a) GW b) SOG c) WP d)GE e) Ours

Figure 5: Assessment among preceding technique



a) Blurred Image b) Deblurred c) Ours

Figure 6: Input and output imagery

Table 1: Estimation of eminence metrics of Image quality Assessment (IQA) and Special Image quality Assessment (SIQA) of Proposed Method

	Methods	**M**	**SD**	**E**	**PSNR**	**RMSE**	**SSIM**	**ODS**	**HCC**	**CG**	**UQI**
ID	ASDS	127.3	76.98	7.911	25.31	19.916	0.694	67.867	0.444	0.465	0.948
	ASDSAR	127.31	76.834	7.919	25.73	19.117	0.716	64.82	0.475	0.422	0.962
	ASDSARNL	127.34	76.66	7.933	26.16	17.816	0.742	58.449	0.531	0.334	0.940
CA	GW	129.6	72.443	7.798	65.038	0.142	0.9983	62.604	0.070	0.054	0.976
	SOG	128.73	72.057	7.794	65.745	0.132	0.9986	62.05	0.096	0.051	0.976
	WP	127.15	76.134	7.923	69.992	0.081	0.9999	58.449	0.477	0.045	0.981
	GE	127.55	76.954	7.931	69.993	0.081	0.9999	58.449	0.978	0.044	0.981
	Ours	127.42	76.812	7.910	**70.002**	**0.081**	**0.9999**	58.449	0.962	0.044	**0.981**

Generally, higher ODS, clearer the improved picture.

e) Universal quality index (UQI)

It mainly assess performance of UQI among original as well as the improved imagery given as

$$Q = \frac{4\sigma_{de}\,\bar{d}\,\bar{e}}{\left(\sigma_d^2+\sigma_e^2\right)\left[\left(\bar{d}\right)^2+\left(\bar{e}\right)^2\right]} \tag{33}$$

Where

$$\bar{d} = \frac{1}{N}\sum_{m=1}^{N} d_m, \quad \bar{e} = \frac{1}{N}\sum_{m=1}^{N} e_m$$

$$\sigma_d^2 = \frac{1}{N-1}\sum_{m=1}^{N}\left(d_i-\bar{d}\right)^2, \quad \sigma_e^2 = \frac{1}{N-1}\sum_{m=1}^{N}\left(e_m-\bar{e}\right)^2$$

$$\sigma_{de} = \frac{1}{N-1}\sum_{m=1}^{N}\left(d_m-\bar{d}\right)\left(e_m-\bar{e}\right)$$

Traditional UQI criterion both uses a picture with high excellence as the reference image. Thus, higher UQI, better the compared image. However, the deblurred imagery is always chosen as of reference imagery, so large UQI do not means that the imagery is of high excellence. The improved representation with best visibility may have smallest UQI.

6. Conclusion

We projected an effective technique namely an underwater turbidity removal scheme with adaptive sparse domain selection (ASDS) methodology significantly improves undersea imagery, and consequently, results of image deblurring. If the deblurred imagery has any lighting problem, then; we used color corrected method using Weighted Gray Edge to find the clear image.

The experimental results shown that the projected means illustrate towards removing turbidity outperforms several state-of-the-art restoration scheme in excellence metrics along with ocular excellence. Finally, we use a color correction scheme to clear the lighting (color) in state-of-the-art images.

Acknowledgment

The authors thank the reviewers for their thorough and helpful remarks.

References

[1]. M. Kocak, Fracer Dal, M, F Cai and Sche, "A focus on recent development and trends in Underwater Imagng," *Mare Tech. Socie. Jorn*, Mar-2008.

[2]. Gian. Lu. F, "Visual inspection of Sea Bottom Structures by an Autonomous Underwater Vehicle," *IEEE Transac Systm., Man, Cybern.* Oct-2001.

[3]. Alberto O, Miqel S, and Gab.O, "A Vision System for an Underwater Cable Tracker," *Machn. Vision. Appln.*, 2002.

[4]. Adria. O and Eman. T, "Detecting Man-Made Objects in Unconstrained Subsea Videos," *Prod. BMV*-2002,

[5]. BA. Level and L. Berg, "Control of Underwater Vehicles in full unsteady flow," *IEEE Jrn. Oceac. Engg*-2009.

[6]. CH. Maze, "In Situ Measurement of Reflectance and Fluorescence spectra to support Hyperspectral Remote Sensing and Marine Biology Research," *IEEE oceanc*-2006.

[7]. Yaac. K and Jeffrey R, "Analysis of Hull remains of the Dor D vessel," *Intn. Jrn. of Nautial Archeo*- 2001.

[8]. BL. McGla, "A Computer model for Underwater Camera System," *Proc of spie*1979.

[9]. J. Jaf, "Computer Modeling and Design of Optimal Underwater Imaging Systems," *IEEE Journ of Oceanic Engi* -1990.

[10]. J. Fun, B. Brya, and J. Heck, "Handbook of Underwater Imaging System Design," TP303 - 1972.

[11]. Timo H, Thom P, F. Chap, and Rob. CT, "A Range-Gated Laser System for Ocean Floor Imaging," *Marin Tech So Jo*-1983.

[12]. John W and Kenn. V, "Point spread functions in Ocean Water: Comparison between Theory and Experiment," *Appd Opti*-1991.

[13]. Tor M and Gre Du, "A statistical learning based method for color correction of Underwater Image", Advn artifi intelli theory-2005

[14]. Chike and Fu"Automatic white balance for digital still camera," *Inf.orm Scien Jrnl- 2006*

[15]. Weng, Ho C, and Fu, "A Novel automatic white balance method for digital still Cameras," IEEE_ Inter .Sym Ckts & Sys- 2015.

[16]. EY La, "Combining grey world & Retinex theory for automatic White Balance in digital photography," Inter Sym Consum Elec, 2005.

[17]. Xial. W, Xiang. Z and Jia.W, "Model-Guided Adaptive Recovery of Compressive Sensing," *Proc Data Compre Confe*, - 2009.

[18]. Juli. M, Franc. B, Jean. P, Guil. S and Andre. Zn, "Non-Local sparse models for Image restoration*," Intern Conf r on Compt Visn*-2009.

[19]. Stef. K, Stanley. O, and Peter. W. J, "Deblurring and Denoising of images by Non-Local Functional," *Multiscl Modl and Simun*- 2005.

[20]. Mattan. P, Michael. E, Hiro. T, and Paymn. M, "Generalizing the Nonlocal-Means to Superresolution Reconstruction," *IEEE Transac. Img Pros*- 2009.

[21]. Riz, Gat and Mar "Color Correction between Gray World and White Patch," Humn vis & Elecr,2002

[22]. Finly and Trez "Shades of grey and colour constancy," Clr & Imag Confr, 2004.

[23]. J Van, T Geve "Edge based colour constancy," proc.IEEE Trans on Img Proces, 2010.

[24]. Gijs, T Gevrs, and Van, "Improving Color constancy by photometric edge weighting," proc. IEEE_Trans. on Patrn Analy & Machine Intelli-2012.

[25]. A.Chrispin Jiji,Vivek M "Underwater Turbidity Removal through lll-posed optimization of sparse modeling," ICPCSI (2017).

[26]. Weish D, Lei Zhg, Gua Sh, Xiaon Wu, "Image Deblurring and Super resolution by Adaptive Sparse Domain Selection and Adaptive Regularization", IEEE Transc on Imge Proc- 2011.

[27]. We Dong, Le Zhang, Guan Shi, Xiao W, "Image Reconstruction with Locally Adaptive Sparsity and Nonlocal Robust Regularization", Sigl Proces: Img Commn -2011.

[28]. Mauo Delb, Pu Mu , Andr Alman "Non –Parametric Sub-Pixel local point spread function estimation," *Img Proces*-2012.

[29]. Yoav. S and Nir. K, "Recovery of Underwater Visibility and Structure by Polarization Analysis," *IEEE Jol of Ocea Engg*, 2005.

[30]. Y Xu, , J Wen, L Fei1, and Zheng Z, "Review of Video and Image Defogging algorithms and related studies on Image Restoration and Enhancement", IEEE open acces Jorn-2016**.**

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 97-104 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Virtual Output Queues Architecture for High Throughput Data Center Nodes

Angelos Kyriakos[1,2], Ioannis Patronas[1,2], Georgios Tzimas[1], Vasileios Kitsakis[1], Dionysios Reisis[*1,2]

[1]*National and Kapodistrian University of Athens, Electronics Lab, Physics Dpt, GR-15784, Zografos Greece*

[2]*Institute for Communication and Computers (ICCS), National Technical University of Athens, Greece*

ARTICLE INFO

Article history:
Received: 25 July, 2018
Accepted: 12 September, 2018
Online: 22 September, 2018

Keywords:
Data Centers
Virtual Output Queues
FPGAs

ABSTRACT

The latest design approach for Data Centers (DCs) follows the direction of exploiting optical switching to connect Top-of-Rack (ToR) switches that serve thousands of data storing and computing devices. A ToR's usual function is the Virtual Output Queues (VOQs), which is the prevalent solution for the head-of-line blocking problem of the DC switches. An effective VOQs architecture improves the DC's performance by reducing the frames communication latency and it is efficient with respect to the implementation cost. The current paper introduces a VOQs architecture for the ToRs of DCs that function with Time Division Multiple Access (TDMA). The proposed VOQ architecture contains a bounded number of queues at each input port supporting the active destinations and forwarding the input Ethernet frames to a shared memory. An efficient mechanism of low latency grants each queue to an active destination. The VOQs constitutes a module of a ToR development, which is based on a commercially available Ethernet switch and two FPGA Xilinx boards, the Virtex VC707 and the Xilinx NetFPGA. The VOQs architecture's implementation and validation took place on the NetFPGA board.

1. Introduction

Data centers are comprised of a large number of Servers running Virtual Machines (VMs) and storage resources, which are installed in racks and communicate via the local data center network. The data centers performance depends on the available computing and data storing capacity, the architecture and the features as well as the performance of the underlying network and the Top-of-Rack (ToR) switches connecting the servers to the data center. A key factor in improving the performance of the ToR switches is the solution of the head-of-line blocking issue that is most often settled by embedding Virtual Output Queues architectures [1]. The performance of the networks depends on their interconnection scheme, which usually adhered to the multi-layer approach, and they were based on the Fat Tree or the folded Clos architectural schemes [2, 3, 4]. These approaches nevertheless, are not efficiently scalable and also, in the cases of data centers with a large number of nodes, lead to the use of a considerable number of switches, cables and transceivers, which increase power consumption.

In an effort to overcome these deficiencies researchers and engineers have introduced data center interconnections including an optical circuit switching as well as an electrical packet switching networks [5, 6, 7]. A notable design is the all optical data center proposed by the Nephele project [8]. The Nephele design adopts the Time Division Multiple Access (TDMA) mode of operation in the optical data center network. Consequently, the transmissions are completed within fixed time segments, namely the slots; each slot is assigned for sending a TDMA frame on a specific path that connects a transmitter node to a receiver node. The Nephele data center network is a Software Defined Network (SDN) and all the arrangements regarding its operation are dictated by a central data center controller. The controller is responsible for generating the TDMA Schedule, which defines which nodes communicate during each time-slot [9]. The first version of the scalable, high capacity Nephele network is able to accommodate up to 1600 Top-of-Rack (ToR) switches and each ToR uses 20 links to connect to the data center optical network.

The overall system topology of the data center network is depicted in Figure 1. The network includes I ($I \leq 20$) parallel planes, each consisting of I ($I \leq 20$) unidirectional optical rings.

*Dionysios Reisis, +30 210 727 6708/6720 & dreisis@phys.uoa.gr

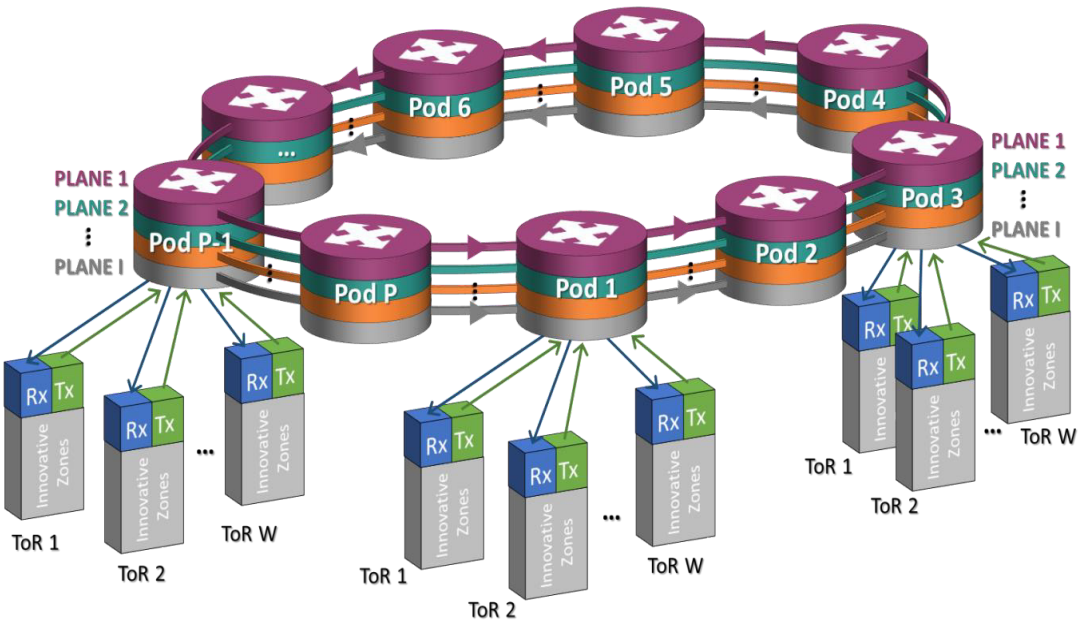


Figure 1: Nephele Data Center Network Architecture

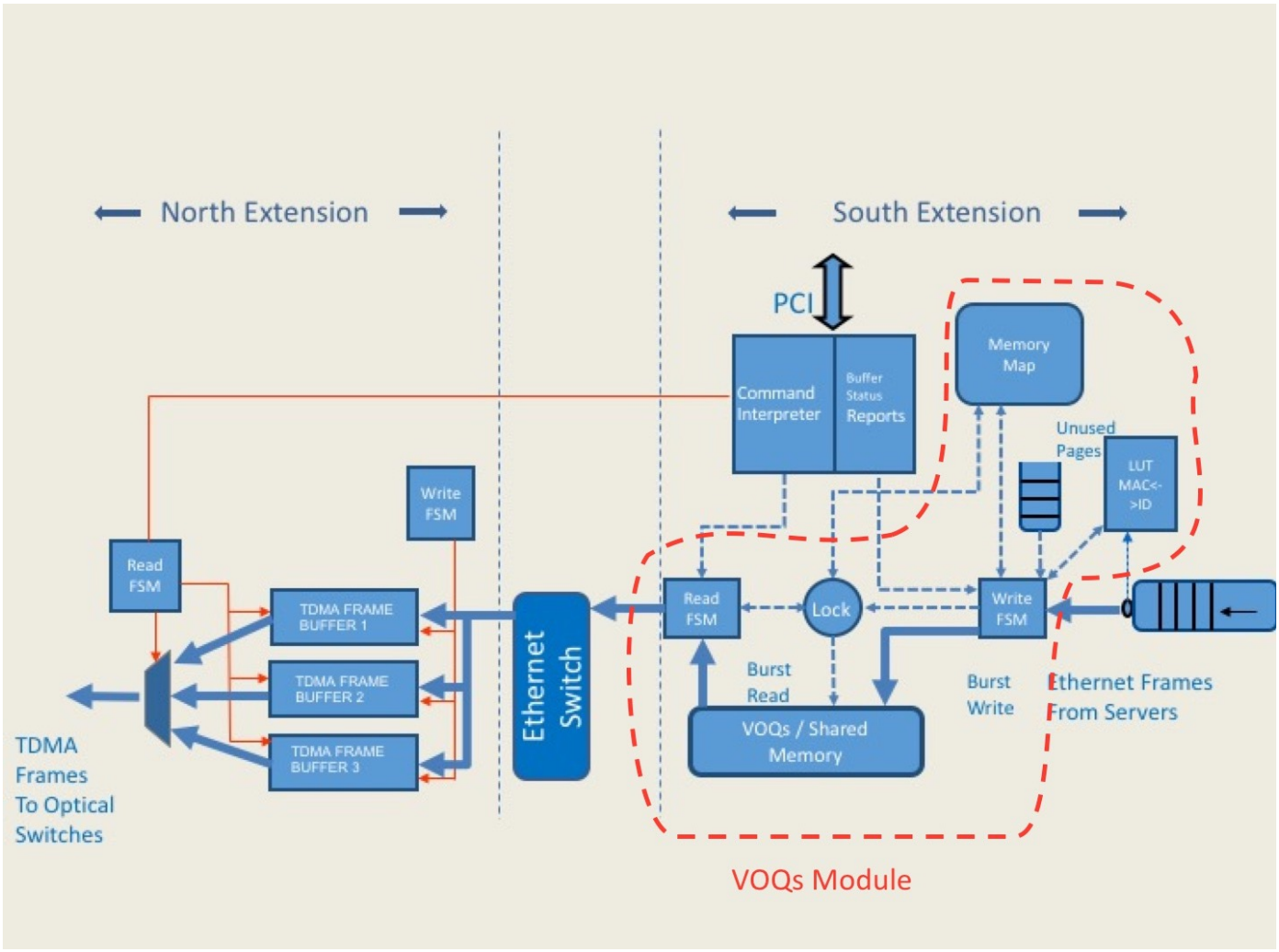


Figure 2: The Nephele Top-of-Rack (ToR) Switch Architecture Overview

The rings interconnect P (P ≤ 20) Points of Delivery (PODs). A POD comprises of I Wavelength Selective Switches (WSS) to connect the I rings, and is connected to W (W ≤ 80) ToRs, through W pod-switches, one for each ToR switch. Each ToR switch has I north ports, such that the *i*th north port is directed to the *i*th POD of each plane (each port is connected to a different POD switch). The south ports of the ToR switch connect the servers, though network interface cards (NICs), with the data center network. The performance of the ToR switch contributes significantly to the operation of the entire data center and it depends on the utilization of its resources as well as on the efficiency of the algorithms and the techniques that it employs. Among the techniques that are critical with respect to the ToR's performance is the handling of the Virtual Output Queues (VOQs). VOQs is an attractive technique for overcoming the head-of-line blocking cases [10].

This paper presents an efficient VOQ organization regarding the resource utilization and the latency needed to assign the incoming Ethernet Frames to the queues of matching destinations. The VOQ architecture introduced in this paper is advantageous due to the following: first, it is efficient with respect to the required implementation area, because it reduces the resources needed to a single shared buffer per output port. This buffer stores all the queues of data that this output port will transmit. Second, the architecture is efficient with respect to the utilization of the shared buffer's bandwidth; this is because it maximizes the throughput utilization of the buffer's interface by utilizing for storing and reading a paging organization, with each page containing a large number of Ethernet frames. Third, the proposed VOQ architecture is scalable, which is an advantage considering the scalability of the entire data center.

The proposed technique achieves the aforementioned goals based on the following ideas. The receiving ethernet Frames with the same destination are collected at the input of the switch into pages of frames. This operation is accomplished by using small sized queues positioned at each input Ethernet port. In the proposed design the number of these small sized queues at the input is bounded by the sum of the connections that are: a) serviced by each input Ethernet port and b) active during a small window of time. The latency is minimized with respect to the time required to associate each input Ethernet Frame to one of the queues. This is accomplished by employing a mechanism that maps each small size queue to one of the active destinations each time an Ethernet Frame arrives at the ToR.

The motivation for designing the proposed VOQs architecture came by the requirements of the ToR included in the Nephele project but it can serve any network, that receives an input of Ethernet frames and particularly those networks which operate under TDMA scheme, are software defined and their nodes may have to overcome the head-of-line blocking. The prototype Nephele ToR switch includes a commercially available Ethernet switch (Mellanox SX1024 [11]) and two Xilinx boards: one Virtex VC707 and one NetFPGA SUME [12]. The implementation and the validation of the VOQs architecture took place on the NetFPGA board.

The paper is organized in five sections. Section II briefly highlights the architecture of the Nephele ToR switch. Section III introduces the architecture and the organization of the VOQs. Section IV presents the details of the FPGA implementation and finally, Section V concludes the paper.

2. The Architecture of the Top-of-Rack (ToR) Switch

The ToR design is a switch and its ports are divided in two sets: a) the south ports, which are 16 10GEthernet ports connecting the ToR with the servers b) the corresponding 16 10Gbps north ports that are connected to the optical data center network. The ToR switch consists of three fundamental blocks. The first is an Ethernet 16×16 switch having all ports as 10GEthernet [11]. The second is the North Extension. It is implemented on an FPGA and its role is: a) the formation of TDMA frames that consist of Ethernet frames and b) to implement the interface of the ToR to the network's optical (POD) switches by using its north ports. The third block is the South Extension. This FPGA based block connects the servers to the ToR. It has increased complexity and its functionality includes: a) the execution of the scheduling commands, b) to be responsible for the communication of the ToR to the data center's control plane, c) to implement the VOQs design and d) to control all the functions of the ToR.

Figure 2 presents the ToR switch's architecture as well as the functional blocks dedicated to the upstream traffic. In the part of the South Extension the figure shows the *LUT MAC-ID* that it assigns a tag to each incoming Ethernet frame. These 11 bits tags will be used within the ToR for addressing the Ethernet frames and saving on the required resources for address bits with respect to the bits required for the MAC addresses of the destinations of the incoming frames. The next action is to forward the Ethernet frames to the *VOQs/Shared Memory* block. This block stores the Ethernet frames in pages. Each page includes a large number of Ethernet frames and its length matches the length of a TDMA frame (also called Nephele frame). All the pages that belong to a destination are arranged in a linked list. The pointers required for keeping the information of each destination's linked are managed by the *Memory Map* block.

The *Command Interpreter* block (Figure 2) is responsible for the translation of the SDN controller commands: it provides to this ToR the destination ToR, which has to receive data in the upcoming TDMA slot. The ToR complies to this command and it retrieves the first page with Ethernet frames that belongs to the linked list associated to the commanded destination and sends this page to the *Ethernet switch*. There is a *Lock* mechanism (Figure 2) that grants either the storing operation of the input Ethernet frames to the shared memory or the reading operation from that memory of the TDMA frames. In more detail, the *Lock* mechanism divides the time into small time windows T_L. Each T_L is dedicated for either writing to the shared buffer or reading from

it. Hence, when the ToR reads from the shared buffer it will continue buffering in the small size queues the incoming traffic from the servers. The length of the T_L is computed at design time to balance: first, the throughput of the shared buffer, which requires long burst transactions for improved performance and second, the need of the ToR operation for writing/reading to/from the shared buffer at close time instances.

The role of the *Ethernet switch* in the upstream direction, is to forward the Ethernet frames to the North Extension and particularly to the buffer of the corresponding destination's north port. In that buffer the Ethernet frames formulate the final TDMA/Nephele frame, to which are also added first, the preamble and second, a word required for each device synchronization. The *Command Interpreter* follows the schedule received from the control plane servers and it specifies (the red control signal of Figure 2) the slot that the ToR will transmit that TDMA frame. For the downstream direction, the Nephele design mandates the Ethernet switch to just forward the frame from the north input port to the corresponding south port. That is, the design complexity of the ToR is mostly related to the upstream path.

The communication of the ToR switch with the control plane is accomplished through the PCI Express interconnection. The PCI Express interface in the proposed architecture is implemented by the use of the Xilinx IP Core for PCIe and RIFFA (Reusable Integration Framework for FPGA Accelerators) [13]. The RIFFA framework consists of an API (Application Programming Interface) and a driver/kernel module for the host PC and IP core for the FPGA, all of which are open-source. The module provided by RIFFA for the FPGA is designed as an extension to the Xilinx core, which handles the physical layer of the PCIe interface.

3. Virtual Output Queues

The proposed VOQ design improves the required hardware resources based on the following concept. During a narrow time window T_B, the ToR switch receives Ethernet frames at its south ports for various destinations in the data center network, which we define as *active destinations*. We consider that for all practical purposes, the number of the active destinations, during T_B, has an upper limit, which can be an outcome of statistical measurements of the network traffic patterns. The active destinations' upper limit is significantly smaller compared to the number of all the possible destinations in the data center. Hence, letting a queue to keep all the incoming Ethernet frames during T_B that have the same active destination and prepare in this queue a burst to be written to the shared buffer, leads to an architecture that includes a set of queues with cardinal number equal to that of the active destinations, while it still keeps the high throughput at the shared buffer.

Considering the above, the VOQs architecture is comprised of: first, the Shared Memory (buffer), second the Memory Map depicted in Figure 2 and third, the VOQs controller. The detailed architecture of the VOQs controller is shown in Figure 3: it is a design of the VOQs controller that includes four (4) active destinations and the corresponding queues, based on a hypothesis that the application asks for four active destination and as shown in Figure 3 there is one queue to support each active destination. In order to define the length of the time window T_B we consider the following facts..

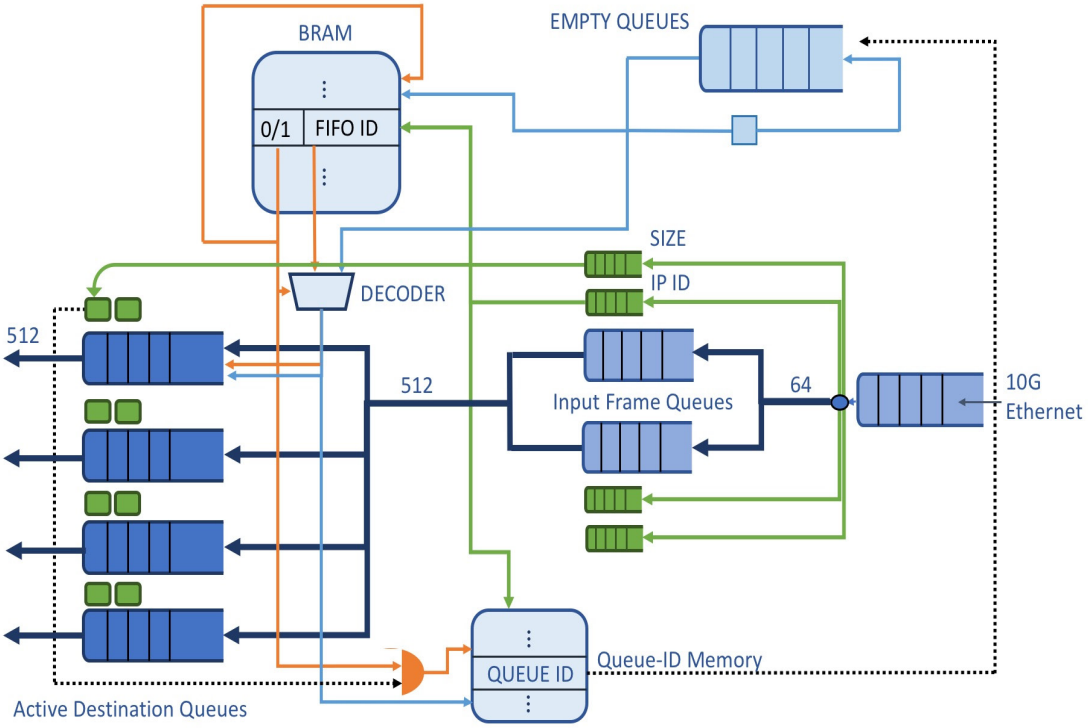


Figure 3: VOQs Controller Architecture Overview

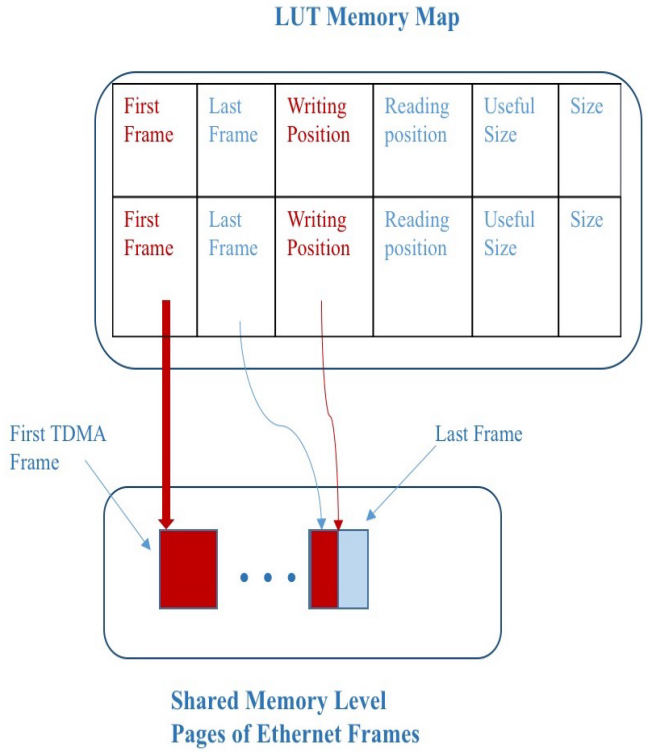
The design of the shared buffer employs a Dynamic Random Access Memory (DRAM) that can reach the considerable throughput of 80 Gbps at its interface; this performance will be feasible if the entire VOQs architecture can operate with burst transactions for reading and writing from/to the shared buffer, thus exploiting the DRAM interface, which requires a minimum burst time t_{mb} depending on the DRAM specifications. The performance of the DRAM organization degrades significantly when the size of the burst size decreases. We note here that, this performance degradation cannot be expressed (defined) as a function of the burst size, e.g. proportional. Therefore, reading and writing from/to a page in the shared buffer (in the linked list assigned to a destination) must be performed in bursts and each burst has to consist of multiple Ethernet frames, in the order of Kbytes. Therefore, we need an architecture of queues able to gather into a single queue all the incoming Ethernet frames that have the same destination; in that queue, the controller will formulate a burst of these Ethernet frames. Finally, it will operate in burst mode to store these frames into the page of the linked list of that destination, which is kept in

We consider the time window T_B and the number of queues for active destinations k to be calculated by the following reasoning. At a clock cycle T_0, given that are available k queues storing Ethernet frames of k different IP, there will be Ethernet frames arriving to at most all of these queues and at the clock cycle T_b that at least one of these queues has completed a burst, and this queue can write the burst to the buffer. Therefore, this queue can formulate another burst either for the IP that it was supporting up to T_b or the queue can be reassigned by the controller to serve another IP. Thus, in this scenario, the worst case is that we have to keep the k queues serving their IPs for as long as no queue has completed a burst: assuming that each queue receives an Ethernet frame in a round robin fashion T_B is at most equal to $k \times t_{mb}$.

According to the above, the efficiency of the VOQs architecture is defined as the maximization of the utilization of the available resources and the DRAM buffer throughput. For this purpose, the design has to: a) include k queues for preparing the bursts, so that each queue prepares a burst that will be stored in an active destination linked list of pages; b) minimize latency and c) minimize the number of the k queues along with their size. The su

cceeding paragraphs describe how we achieve the above goals and they describe in detail the operations of the VOQs Controller as well as its functional blocks and components.

The ToR switch is connected with 10G Ethernet to the servers through its south ports. First, the Ethernet Frames that arrive from the servers at the rate of 10G are buffered in the port queue of the 10G Ethernet module and then are forwarded and buffered to the two *Input Frame Queues* (Figure 3) in the following way: we start counting the incoming frames and depending on the arrival sequence the odd numbered incoming Ethernet Frames are stored in the first *Input Frame Queue* (the upper queue on Figure 3) and the even Ethernet Frames to the second queue. This dual queue architecture gives us the necessary time in order to perform in real-time the two following operations on the Ethernet Frames: while we store a frame in one of the *Input Frame Queues*, we calculate its size and extract its destination's IP, which then are stored to two queues of significantly lesser size, the *IP ID* queue and the *SIZE* queue, which are positioned close to each *Input Frame Queue* in the design of Figure 3.



Each frame's IP stored in the *Input Frame Queues* is passed as input (address) to a LUT, named *BRAM* in Figure 3. The LUT will specify (will give as output) the id of an *Active Destination Queue* (on Figure 3 we shown an example design with four queues): in the specified *Active Destination Queue* we will buffer all the Ethernet Frames with the current active destination IP, in order to form a burst that it will be stored into the linked list of pages of that destination in the DRAM buffer. Apart the id of the *Active Destination Queue* in that *BRAM* location is also stored a flag (0/1). When the flag is equal to "1", it specifies the case in which the *Active Destination Queue* id (stored in the LUT) is granted to the active destination IP. Alternatively, the case when the flag equals to "0" indicates that the frame's destination IP is not yet served by any of the *Active Destination Queues* and hence, the controller has to assign an *Active Destination Queue* to this IP. Now, we consider the case of an Ethernet frame arriving at the ToR and its IP address does not correspond to any of the *Active Destination Queues.* If we have correctly calculated (during the design of the ToR) the minimum required number of the *Active Destination Queues* that it is sufficient to serve the application demands, the VOQs controller will have an empty *Active Destination Queue* available for assignment to a newly arrived

Ethernet frame that requests an *Active Destination Queue* to buffer the following frames with the same IP destination. All the id (numbers) of the unused *Active Destination Queues* are buffered in the queue named *Empty Queues* in Figure 3. At the same clock cycle that we read from the *BRAM* the id of the *Active Destination Queue* that serves the frame's IP along with the "1/0" (assigned to a queue or not) flag, we also read the first empty queue id from the *Empty Queues*. The multiplexer shown at Figure 3 bellow the BRAM is controlled by the flag in order to select: a) the *BRAM* output when the flag equals "1" and b) the *Empty Queues* output if the flag is "0". In the first case where we will use the *BRAM* output, the empty queue id that was just extracted from *Empty Queues* will be returned back in the *Empty Queues,* since it was not used. The above design minimizes the latency for the assignment of an active queue to the new destination.

We have to mention that in order to exploit the high throughput of the DRAM interface, we have to write the Ethernet Frames in the shared buffer as a burst of contiguous words of a significant length (512 bits in the example implementation of the proposed architecture). We note here that, in a writing burst of Ethernet frames the last 512-bit word might not be completely filled with Ethernet frames payload and for completing the burst we add 0xFF as padding. The simple padding provides the advantage of simplifying the control and it reduces the latency at the cost of the dummy data overhead in many pages in the shared buffer. This padding overhead becomes larger for small Ethernet frames and it is reduced significantly in the case of full Ethernet frames. Note here that, when it's time to transmit a TDMA frame the shared memory will provide us with a page: we must be informed regarding the exact number of the useful data in this page in order to remove the padding. For this purpose, we store in the header of each page the useful size along with the actual page size, which is the overall sum of the useful size and the size of the padding stored in the shared buffer.

A small size dual port memory shown in Figure 3. as *Queue-ID Memory*, stores the IP that it is currently served by each *Active Destination Queue*. Each address *X* of the *Queue-ID Memory* corresponds to the *Active Destination Queue* with id *X*. The data at that address *X* of the *Queue-ID Memory* is the destination's IP that is accommodated by this *Active Destination Queue*. When it is the first time that an Ethernet Frame is stored in an empty *Active Destination Queue* the id of this queue is used as the address to the *Queue-ID Memory*, and in that address, we store the frame's IP. During the whole time that this *Active Destination Queue* serves the IP, the *Queue-ID Memory* keeps the IP in that address. Only when an *Active Destination Queue* is left with all its data forwarded to the shared buffer, we will: first, erase the contents of the served destination in the *BRAM* by acquiring the address (IP) from the *Queue-ID Memory* and second, write the queue id to the *Empty Queues* to refresh the *Active Destination Queues* that are vacant and they can be granted to another destination IP. Consequently, the location in the *Queue-ID Memory* will be overwritten by the new IP, which will be served by the corresponding *Active Destination Queue*.

The proposed design minimizes the time required to perform all the previously mentioned operations with respect to clock cycles. The architecture can achieve the time minimization due to the parallelization of the operations and as a result, the VOQ architecture diminishes the latency of each stage. Consequently, the *Active Destination Queues* can be as many as the application dictates as upper bound. Moreover, the length of each queue doesn't need to grow beyond the size of the burst that it is specified by the DRAM controller for reaching its maximum throughput.

The block called *Memory Map* stores all the information related to each linked list in the shared buffer associated to each destination IP. The memory map entries are shown in Figure 4, 5 in two working examples. Each entry of the *Memory Map* block has the following pointers: one at the address of the first page of the list noting from what page we are currently reading data to transmit; one to the last page, required to inform the VOQs that this is the page, which currently stores all the Ethernet frames for the associated destination; one for the "next to write" address of the last page (writing position in Figure 4), one for the "next to read" address of the first page (reading position in Figure 5). Moreover, the *Memory Map* entry provides the exact number of useful data in the page: this information is used to compute the total volume of data of the Ethernet frames with or without out the padding.

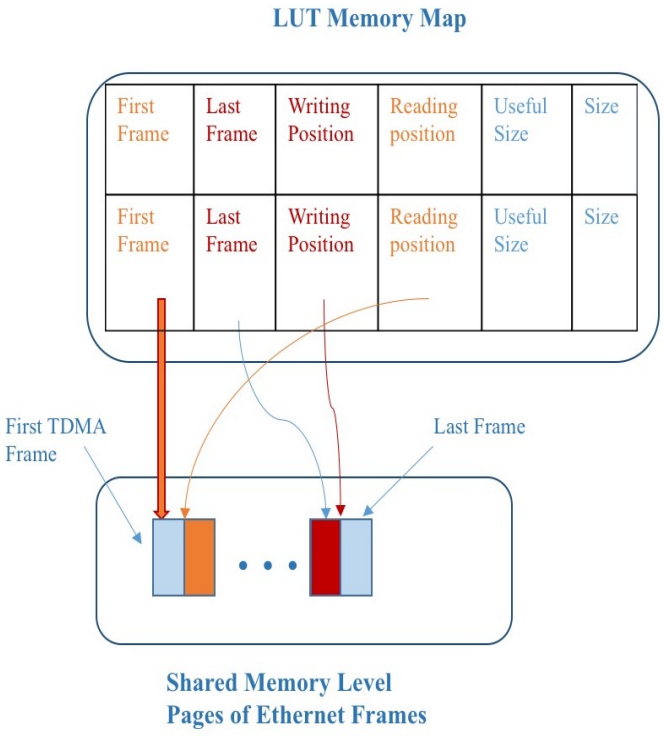


Figure 5: Memory Map Organization Concurrent Write & Read Operations

The pointers at each *Memory Map* block location are refreshed during each burst write/read transaction. Thus, at the beginning of a write/read operation to/from the DRAM buffer we know the exact number of the data (bytes) that will be written/read. We operate the linked list of pages as a queue since we always transmit the head page of the list. The memory map architecture is able to concurrently write and read from the same linked list of pages as shown in Figure 5.

A noteworthy advantage of the novel VOQ technique is the scalability of the architecture, which can be easily configured to accommodate various numbers of Active Destination Queues, the size of the DRAM shared buffer and the size of the *Memory Map* block. The pointers and the size of the linked list of pages for each destination are stored in block rams (BRAM) in the FPGA. The required size of the BRAMs is proportional to: first, the DRAM memory size, and second the number of destinations in the data center network. In the case that the size of the mapping information is relatively large and it constrains the designer of implementing the Memory Map on the FPGA internal BRAM memory, the proposed architecture of the Memory Map block can be implemented on an external Static Random Access Memory (SRAM).

4. FPGA Implementation Details

We have realized an example VOQs design with four (4) active destinations (k = 4 is adequate for most applications in accordance with our T_B and k calculations). The development of the example implementation was made on the NetFPGA SUME board using the Xilinx Vivado development tool. The design includes 3 Intellectual Property (IP) hardware Cores from Xilinx: a) 10GbE Subsystem, which includes the MAC and the 10GbE PCS/PMA b) Integrated Block for PCI Express c) Memory Interface Generator (MIG) for the shared DRAM buffer. The NetFPGA board receives the scheduling commands from the host desktop PC, which is running Linux and communicates with the data center's controller, which runs on a different PC in the same local network.

Table 1: FPGA Resources

Slice LUTs	4194
Slice Registers	2415
Slices	1888
Logic LUTs	2639
Memory LUTs	1285
Flip flop pairs LUTs	4848
Block RAMs	62

The resources occupied in the NetFPGA SUME for the VOQs Controller are presented in the Table 1, reported by the Vivado tool. The input small sized queues are all performing at 156 Mhz clock and use 64 bits word length, in order to comply with the 10G Ethernet physical layer standard. The Active Destination Queues and memories alongside of them in our implementation are performing at 200 MHz with 512-word length.

5. Conclusion

The current paper presented a VOQs architecture, which is efficient with respect to latency and the hardware resources and it supports a ToR switch that is adaptable to any data center network operating under the TDMA scheme. The most noteworthy novelty of the proposed VOQs architecture is the efficient use of a single large shared buffer, the performance of which is fully exploited. The VOQ organization is based on the notion of *Active Destination Queues* that leads to maximize the utilization of the shared buffer and reduces significantly the required number of the *Active Destination Queues* to the number of the connections that are active during a narrow time window. Moreover, the management/control of the *Active Destination Queues* is efficient due to the minimum latency that induces to the operation of the ToR switch. Furthermore, the proposed architecture is scalable with respect to the number (k) of the *Active Destination Queues,* the scale of the data center network (number of destinations), the shared buffer size and the Ethernet protocol (Ethernet type/Frame size).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Kyriakos, I. Patronas, G. Tzimas, V. Kitsakis and D. Reisis, "Realizing virtual output queues in high throughput data center nodes," 2017 Panhellenic Conference on Electronics and Telecommunications (PACET), Xanthi, 2017, pp. 1-4. doi: 10.1109/PACET.2017.8259971

[2] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," SIGCOMM Comput. Commun. Rev., vol. 38, no. 4, pp. 63–74, Aug. 2008. [Online]. Available: http://doi.acm.org/10.1145/1402946.1402967

[3] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: A scalable and flexible data center network," in Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, ser. SIGCOMM '09. New York, NY, USA: ACM, 2009, pp. 51–62. [Online]. Available: http://doi.acm.org/10.1145/1592568.1592576

[4] N. Farrington, E. Rubow, and A. Vahdat, "Data center switch architecture in the age of merchant silicon", in Proceedings of the 2009 17th IEEE Symposium on High Performance Interconnects, ser. HOTI '09, 2009, pp. 93–102.

[5] H. H. Bazzaz, M. Tewari, G. Wang, G. Porter, T. S. E. Ng, D. G. Andersen, M. Kaminsky, M. A. Kozuch, and A. Vahdat, "Switching the optical divide: Fundamental challenges for hybrid electrical/optical datacenter networks," in Proceedings of the 2Nd ACM Symposium on Cloud Computing, ser. SOCC '11. New York, NY, USA: ACM, 2011, pp. 30:1–30:8. [Online]. Available: http://doi.acm.org/10.1145/2038916.2038946

[6] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in Proceedings of the ACM SIGCOMM 2010 Conference, ser. SIGCOMM '10. New York, NY, USA: ACM, 2010, pp. 339–350. [Online]. Available: http://doi.acm.org/10.1145/1851182.1851223

[7] K. Tokas, C. Spatharakis, I. Kanakis, N. Iliadis, P. Bakopoulos, H. Avramopoulos, I. Patronas, N. Liakopoulos, and D. Reisis, "A scalable optically-switched datacenter network with multicasting," in 2016 European Conference on Networks and Communications (EuCNC), June 2016, pp. 265–270.

[8] P. Bakopoulos, K. Christodoulopoulos, G. Landi, M. Aziz, E. Zahavi, D. Gallico, R. Pitwon, K. Tokas, I. Patronas, M. Capitani, C. Spatharakis, K. Yiannopoulos, K. Wang, K. Kontodimas, I. Lazarou, P. Wieder, D. Reisis, E.

Varvarigos, M. Biancani, H. Avramopoulos, “NEPHELE: an end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud datacenters”, IEEE Communications Magazine, 2018

[9] K. Christodoulopoulos, K. Kontodimas, K. Yiannopoulos, E. Varvarigos, "Bandwidth Allocation in the NEPHELE Hybrid Optical Interconnect", 2016 18th International Conference on Transparent Optical Networks (ICTON), July 2016.

[10] Pedro Yébenes, German Maglione-Mathey, Jesus Escudero Sahuquillo, Pedro J. García, Francisco J. Quiles, “Modeling a switch architecture with virtual output queues and virtual channels in HPC-systems simulators”, 2016 International Conference on High Performance Computing & Simulation (HPCS), Innsbruck, Austria, July 2016.

[11] Mellanox Technologies, “SX1024: The Ideal Multi-Purpose Top-of-Rack Switch”, White Paper, May 2013. Available: https://www.mellanox.com/pdf/whitepapers/SX1024-The-Ideal-Multipurpose-TOR-Switch.pdf

[12] N. Zilberman, Y. Audzevich, G. A. Covington, and A. W. Moore, “Netfpga sume: Toward 100 gbps as research commodity,” IEEE Micro, vol. 34, no. 5, pp. 32–41, Sept 2014.

[13] Matthew Jacobsen, Dustin Richmond, Matthew Hogains, and Ryan Kastner. 2015 RIFFA 2.1: A Reusable Integration Framework for FPGA Accelerators. ACM Trans. Reconfigurable Technol. Syst. 8, 4, Article 22 (September 2015), 23 pages. Available: http://dx.doi.org/10.1145/2815631

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 99-111 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Extendable Board Game to Facilitate Learning in Supply Chain Management

Linda William*[1], Za'Aba Bin Abdul Rahim[1], Robert de Souza[1], Eko Nugroho[2,3,4], Rio Fredericco[2]

[1] *The Logistics Institute Asia Pacific, National University of Singapore, 21 Heng Mui Keng Terrace #04-01, Singapore 119613*
[2] *Kummara Studio, Jl. Sidomukti no. T1, Bandung 40123 Indonesia*
[3] *Center for Research on Information & Communication Technology, Institute Technology Bandung, Jl. Ganesha No.10, Lb. Siliwangi, Coblong, Kota Bandung, Jawa Barat 40132, Indonesia*
[4] *GameLab, Computer Science Program, Bina Nusantara University, Joseph Wibowo Center Campus, Jl. Hang Lekir I No. 6 Senayan, Jakarta 10270, Indonesia*

ARTICLE INFO

Article history:
Received: 14 June, 2018
Accepted: 08 July, 2018
Online: 23 July, 2018

Keywords:
Game-based learning
Facilitate learning
Supply chain management
Board game
Expandable framework

ABSTRACT

Game-based learning has been introduced as an interactive tool to facilitate learning and training processes in various fields, including supply chain management (SCM). Most of these games are specifically designed to focus on certain scenarios and concepts. For example, the original beer distribution game focuses on a single product supply chain without considering capacity and process reliability into account. This creates challenges for extending the games to cover other concepts in SCM. To tackle this problem, we propose a board game, titled ThinkLog, as a face-to-face extendable framework to facilitate learning in SCM. It can be extended to generate different scenarios for various concepts in SCM without changing the basic game structure. Using this principle, we have extended the basic version of ThinkLog to two other scenarios, namely: humanitarian logistics and urban logistics, by simply modifying the rule of the game. Each scenario would have different learning objectives embedded in the gameplay. The game is also complemented with a computer-based application (digital application) to enhance the overall learning experience and collect relevant data (data gathering) during a game session. These three scenarios have been evaluated on four-interactive sessions with government officials and policy- makers in Indonesia. Each session has been consistent in its acceptance of the game as a tool to facilitate learning in SCM, regardless of the scenarios that we played. Our learning objective evaluation also shows that the game is effective in deepening the players' understanding of SCM concepts.

1. Introduction

This paper is an extension of work originally presented in 2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE) [1]. It focuses on developing and evaluating an extendable board game as a game-based learning framework to facilitate teaching and learning in Supply Chain Management (SCM).

Game-based learning has been introduced to improve training activities and initiatives by stimulating learning engagement and motivation through role playing and repeatability for the players to learn new skills and concept voluntarily [2, 3, 4, 5, 6]. It is a subset of serious games which includes the use of any medium of games (e.g. board games, card games, sports or digital games) for learning purposes [7]. It has been implemented in various fields, such as healthcare [8, 9, 10], military applications [11], city planning [12] and supply chain management [13, 14, 15, 16, 17].

Previous studies have identified the benefits of these type of games in a variety of critical-contexts [3]. It includes enhancing engagement and motivation, improving self-monitoring, recognizing and solving problems, improving decision making, and developing social skills for collaboration and negotiation [3,

*Corresponding Author: Linda William, The Logistics Institute Asia Pacific, National University of Singapore, 21 Heng Mui Keng Terrace #04-01, Singapore 119613, Telp: +65-6516 4281, Email: Lindawati.tliap@nus.edus.ag

18, 19]. It would advance the players' understanding and learning in a scientific knowledge [20]. By using games, players become more involved in solving multiple tasks or challenges. Games provide a hand-on experience where the players need to plan appropriate strategies based on their knowledge and skills, implement these strategies in the experimented environment and respond to the consequences of their strategies. It would promote players' motivation and knowledge transfers in a subsequent realistic context [20, 21].

However, most of the game-based learning in SCM focus on one or a limited number of scenarios. For example, the original beer distribution game is designed for a pure retail distribution game without taking product conversion, capacity and process reliability into account [22] or the blood supply chain only focuses on perishable goods [23]. It cannot be extended (or difficult to be extended) to cover different SCM concepts without changing the game structure or game components. Hence, it is limiting the implementation of the games.

Recognizing this problem, we propose a board game, titled ThinkLog, as an extendable face-to-face learning framework for SCM concept. It uses role-playing and simulation to stimulate discussion and interaction between the players. It creates a safe environment for the players to share their thoughts and practice their new skills without the threat of real-world consequences. ThinkLog can be extended to generate different scenarios for various concepts in SCM without changing the basic game structure. To demonstrate its capability, we have extended the basic scenario of ThinkLog [1] to cover two more complex SCM concepts, namely: humanitarian logistics and urban logistics. These new concepts can be embedded in the game by modifying the rules of the game.

We also develop a digital companion application to complement our board game. Companion apps have been used in several games, such as Dungeon Master's Assistance for Dungeons and Dragon game [24] and smart play system for Yes or Know board game [25], to guide the players during the game and provides a simple calculation for fastening the game processes. Our companion app aims to enhance the overall learning experience and collects the relevant data (data gathering) during a game session. The data collected can be analyzed further to understand the players' behavior and game experience. Using this insight, we would be able to adjust and customize the game to fit with individual learning behavior.

To evaluate the effectiveness of ThinkLog, we conducted four interactive sessions with government officials and policy makers in Indonesia using different scenarios. From those sessions, we learnt that the game (with or without the companion app) is highly accepted by the players. The players were also able to grasp the learning objectives of the game easily. Other than these four sessions, we also conducted a game session with master students in the National University of Singapore, as part of their SCM course, to evaluate the game in a larger group setting. The game acceptance and learning objective evaluation of the players resonates throughout the game session.

The remaining of the paper will be organized as follows: Section 2 presents the literature review and section 3 presents the proposed board game. Section 3 includes the game design, components, scenarios and its companion app. Section 4 discusses the evaluation method to measure the effectiveness of ThinkLog in facilitating learning in SCM. Section 5 reviews ThinkLog adjustment for adoption in classroom with large group setting and Section 6 presents the conclusions and future research direction.

2. Literature Review

In this section, relevant literature in the areas of board games and games for SCM are reviewed.

2.1. Board Games

Board games have been used as entertainment for centuries, starting from the ancient games such as *Go* and *Backgammon* that are still used today [26]. Generally, board games are played using an artwork [27] which incdice the main board, balls, dices, and cards. It has a set of rules that need to be obeyed by all the players. Points or rewards are given according to the rules. The players have a flexibility to change the rules based on the consensus from all the players [25]. They would take turns to play and receive feedbacks immediately. The actions and feedbacks are all apparent to all the players. The players would also be able to discuss with each other openly.

Although it has not been explored extensively as digital games, board games show promising opportunities to facilitate learning and training [28, 29, 30]. It has been proven to be as effective as digital games to facilitate learning [31]. As part of game-based learning, board games are able to enhance the players' engagement and motivate them to learn new skills, including social skills, to get potential rewards from the game through interaction and discussion with other players. It can stimulate the players' ability to think and solve problems and challenges. The nature of board games also implies transparencies and interactions as the core mechanism of the game. Hence, these board games are suitable for collaborative games where all players need to work together as a team to win the game [29] or other similar contexts which highly depend on real human interaction [28].

Board games have been successfully implemented to support learning and training in healthcare [32]. It shows a significant increment of knowledge on nutrition labeling [31], dengue hemorrhagic fever [32], healthy eating [33], chronic disease and risk factors [34], tooth morphology [35], pediatric [36], and cancer treatment [37]. Other than healthcare, board games have also been implemented in urban planning [38], agriculture [39], and software design [40]. The Beer Distribution Game, a well-known game in SCM, is also available on board game version [13, 41].

2.2. Games for SCM

A number of games have been introduced to facilitate learning of Supply Chain Management (SCM) concepts [6], among which are the famous Beer Distribution Game [13, 42, 43], Blood Supply Chain Game [23, 44], Mortgage Service Game [15] and Distributor Game [16, 17].

The beer distribution game is the most well-known game in SCM and part of many SCM curriculum since it was being developed by MIT in the 1960s [43]. There are several variants of the beer distribution game in the literature, these includes the stationary beer game [14], computer simulated beer game (e.g. [45, 46]) and online beer game [42]. This game aims to introduce the

basic concepts of the bullwhip effect and the benefits of information sharing and lead-time reduction. It simulates a supply chain for a single product involving a four-level supply chain composed of a retailer, a wholesaler, a distributor and a manufacturer. Each player takes one role and for each turn, he/she follows the same set of activities, that is to fulfil customer demand, order from supplier, and manage inventories. The goals of this game are to minimize the total holding and backorder costs incurred in the entire supply chain.

Blood Supply Chain Game demonstrates supply and demand balance in healthcare perishable products with limited shelf life and production. It simulates the simplified supply chain of blood units from donors to patients based on a real case study modelling the UK blood supply chain [23]. This game is played from the perspective of the distributor which is the middle-player in the blood supply chain network. It has different game options and parameters to determine the probability of blood supply and demand and fulfillment policy. Each player aims to maximize the number of order fulfillment to the hospitals by considering the limited stock.

Different from beer distribution game and blood supply chain game that focus on a supply chain with finished goods inventory, Mortgage Service Game is intended to teach service-oriented SCM concepts [15] that typically do not hold any inventory and can only manage backlogs through capacity adjustment (such as number of employee's adjustment). It aims to introduce the bullwhip effect resulting from information and capacity adjustment lags and impact of end-user demand information in the service industry. This game represents a simplified mortgage approval process that has four stages, namely: initial processing, credit checking, surveying and title checking. Each player or team of players controls a particular stage and able to adjust its capacity by hiring or firing employees to check all mortgage applications in its backlog for each turn. Depending on the setup, the other stages may be managed by artificial agents or real players. The game also introduces lags on the capacity adjustment processes. The goal of this game is to minimize the total cost for the entire supply chain, resulting from employee salaries and service delays (certain fine or backlog fee would be added to the total cost).

Lastly, distributor game, is operating at a continuous clock with ongoing events and responses to individual decision to mimic the real-time global supply chain [17]. It is different from the other three games where the players need to wait for their turns to make their actions. Distributor game is a specific instance of Global Supply Chain Game (GSCG). It aims to educate the players on critical leadership skills for global supply chain in real time. These skills include strategic leadership, operations management, financial management and information technology.

Table 1. SCM Game Comparison

Category	Beer Distribution Game	Blood Supply Chain Game	Mortgage Service Game	Distributor Game	ThinkLog
Game world					
Type	Role-playing simulation	Role-playing simulation	Role-playing simulation	Role-playing simulation	Role-playing simulation
Platform	Board, Digital, Online	Board, Digital (Excel-based)	Digital	Online	Board
Technical Features	Multi players with four roles; Players wait for their turn to make their actions	Single player from the perspective of distributor	Single player/multi players; Players wait for their turn to make their actions	Single player; Players do not need to wait for their turn to make their actions	Multi players; Players wait for their turn to make their actions
Game Event					
Subject or content areas	Industrial production and distribution system	Supply and demand balance for perishable goods	Service-oriented SCM	Global supply chain	Supply Chain Management
Scenario	One scenario; Single product in a four level supply chain	One scenario; Blood supply chain	One scenario; Mortgage approval process that has four stages	One scenario; A four-level global supply chain	Expendable to different scenarios with different gameplay
Learning objective	Introduce the basic concepts of the bullwhip effect and the benefits of information sharing and lead-time reduction	To improve understanding of complex principles of supply chain for perishable goods	Introduce bullwhip effect resulting from information and capacity adjustment lags and impact of end-user demand information in the service industry	To learn leadership skills that are critical to manage real-time global supply chain.	Different scenarios would have different learning objectives

All players take a role as a distributor in a distribution center in a four level supply chain. Other levels, including competing distributors are played by artificial agents to create a complex and dynamic global environment.

The summary and comparison of these four games and our proposed board game are presented in Table 1. The comparison is based on game categories adapted from [47]. Other than these four games, there are other games designed to illustrate different SCM concepts (see, for example, Lean Leap Logistics Game [48] and The Chain Game [49]).

3. ThinkLog: An Expandable Learning Framework

This section would describe ThinkLog as a special board game designed to enable the extension in the scenarios without changing the basic game structure. The game design, components, scenarios and companion app are reviewed.

3.1. Game Design

As an interactive board game for game-based learning, ThinkLog is carefully designed to balance the entertainment components and the pedagogy (i.e. learning objectives and outcomes). The game needs to have a clear objective, derived from specific SCM concepts, research outcomes and case studies. It also needs to have a clear pedagogy to influence learning and motivate the players to gain specific knowledge and skills related to the learning objectives. However, the game would also need to maintain the enjoyment and fun components, such as fantasy [50, 51], challenge [52, 53], choice [54], mechanism and playability [55]. It enables the players to engage and immerse into the game which would eventually motives the players to achieve new skills and understands new concepts voluntarily in a greater speed.

To balance these factors, ThinkLog is developed by following the game base learning framework [56], as illustrated in Figure 1. The framework has been adapted to include board game elements, such as interaction between the players. For each scenario, specific concepts are translated to the learning objectives and then converted into the specific scenario with clear goals for the players.

The entertainment components are captured by the game characteristics that allow players to be involved and engaged in the game. The characteristics can be categorized into 6 criteria, namely: challenge [52, 53], choice [54], competition, fantasy [50, 51], goal, and rule [57]. The summary of these characteristics is shown in Table 2.

3.2. Game Components

There are four main components of ThinkLog, namely: main board, demand cards, gameplay/rules and game master. These are essential components to play ThinkLog regardless of the scenarios played.

3.2.1. Main board

ThinkLog's main board is a symmetrical map to visualize actions and responses during the game. It is a center of the whole activities in the game. It is also the base for all the engagements and twists during the game. By design, the board serves as a platform that gives players a (pseudo) control of the entire game environment. This sense of control is important to keep the player focus and bring more excitement to the game. The board also shows the player' actions and (action) feedbacks.

To represent the supply chain network and logistics activities, several nodes for the players' logistics facilities or other logistics assets are included. Depending on the game scenarios, the players may be able to determine their facility locations. Logistics activities will occur on the map between these facilities. Common disruptions such as traffic congestions, flooding and facilities' breakdown can be shown in the main board to introduce risk and uncertainty aspects of the supply chain.

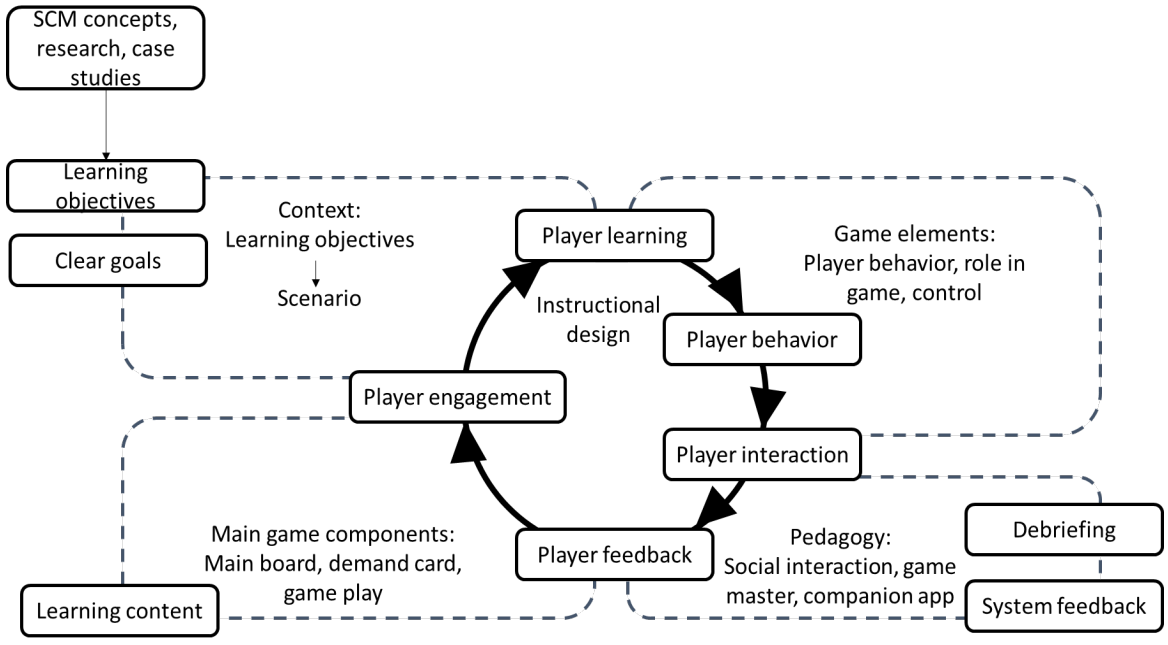


Figure 1. ThinkLog Game Design

Table 2. ThinkLog Characteristics

Challenge	• The difficulty level is determined by the lead time and demand level. • Various disruptions (such as traffic congestions, flooding and facilities' breakdown) would be added.
Choice	• Depending on the scenarios, the players may need to decide the locations of their facilities and consolidate their orders. • Players need to select their vehicle's delivery route. • Players need to decide the number of goods to ship to the next level node in the supply chain and the number of goods to order from the previous level node in the supply chain
Competition	• Depending on the scenarios, the players may compete with the other players to win the game. • The players may compete to get the highest performance (in term of money or other points)
Fantasy	The game uses the main board and other components (i.e. vehicles) to help the players imagine the logistics environment.
Goal (Game Objective)	The general goal is to create an efficient supply chain network.
Rule	• Each player needs to wait for his/her turns to take actions. • Interaction between players and game master are encouraged.

There are two variations of the main board depending on the point or node locations as shown in Figure 2. The first variation, named basic main board, has eight facility locations, while the second variation, named flexible main board, has several alternative locations where the players can choose to locate their facilities or assets.

3.2.2. Demand cards

During the game, players will be introduced to "Demand Cards", as shown in Figure 3. It represents the market demands that need to be fulfilled during the game and players can either work together or individually to do so. The demand values are dependent on the scenarios; it can be randomly generated or based on a certain distribution to mimic real demands for a particular good in the particular scenario.

3.2.3. Gameplay

The gameplay core mechanic in ThinkLog is point to point delivery where the players need to deliver a certain amounts of goods to a certain point. Players main objective is to develop a strategy for efficient and effective deliveries which include managing their delivery timing, amounts, and resupplying their stocks. It will determine the result of the game and players' performance in the game.

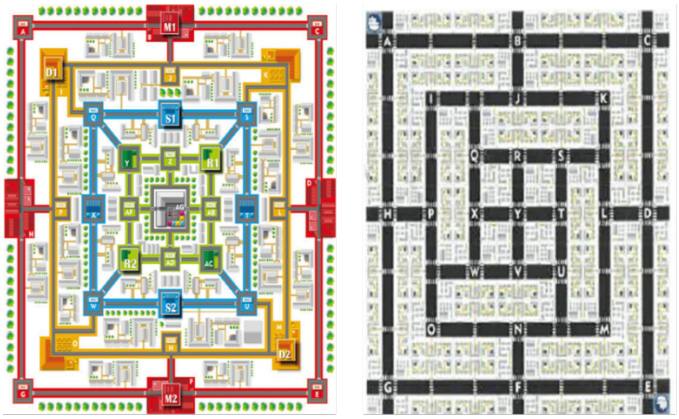


Figure 2. ThinkLog Main Board



Figure 3. ThinkLog Demand Cards

To develop their delivery strategies, the players must consider the following criteria: locations, management cost, facilities and vehicle constraints, and supply availability. Tweaking these criteria would create different situations in the game and eventually resulting in different scenarios. There are 3 main activities for the players within a round of the game as follows.

1. **Ship**. The players send the exact amount of goods ordered by the next level node in the supply chain (i.e. the player's customer). If the players do not have enough goods and cannot fulfil all the order, they will get a backlog token.
2. **Pay.** The players pay inventory cost and backlog cost (if any).
3. **Order**. The players give order to the previous level in the supply chain (i.e. the player's supplier) and pay the price.

The 3 main activities can be re-arranged and adjusted if needed in specific scenarios.

3.2.4. Game Master

A game master guides the overall gameplay of the game session and ensures players understand and follow the rules accordingly. Depending on the number of players and the game master's availability, the game session may be led by one main game master and several supporting/group game masters. For each game session, the game master would need to do the following:

1. **Briefing**. The game master initiates a discussion highlighting on: 1) the concept of the game and its relations with the logistics/supply chain topic; and 2) the overview of the game rules and objectives to guide the players on how to play the game.

2. **Playing**. The game master would be involved in the game, by:

    - Observing the group dynamics.
    - Providing some additional information/motivation to make sure all participants actively involved in the session.
    - Intervening the game in case there are some (unwanted) conflicts. The game master may take necessary action to minimize any possible risk of distraction.

3. **Debriefing**. The game master conducts a discussion after the game session. The game master would motivate the players to share their experience and highlighting some learning points that they found during the playing phase. Later the game master summarizes those findings by emphasizing on main learning points of the game.

Other than these four main components, there are other components for ThinkLog game. It includes player board, truck token, facility token, good token, congestion token, paper money and order form. Some scenarios may need additional components. The use of these components is also depending on the scenarios played. Several components can be replaced by a feature in the companion app. For example, paper money can be replaced by virtual money in the companion app.

3.3. ThinkLog Scenarios

There are 3 scenarios of ThinkLog that have been developed, namely: basic, humanitarian logistics and urban logistics. The two later scenarios are the extensions of the basic scenario. Each scenario is designed to accommodate a certain theme and focuses on different learning objectives. The basic scenario focuses on information flow and coordination challenge in supply-chain and logistics sector. At the same time, it provides an overview of risk management in supply chain and the bullwhip effect. The second scenario focuses on introducing the Multi Criteria Decision Making (MCDM) procedure [58, 59] and its implementation in a simplified humanitarian logistics setting. And the third scenario focuses on important aspects of facility management in the context of last mile delivery. A summary of these scenarios is presented in Table 3.

3.3.1. Basic Scenario

Similar to the Beer Distribution Game [13, 42, 43], each player would choose one out of four available roles, namely: manufacturer, distributor, wholesaler, or retailer. Each role represents one level in the overall supply chain network. Each player or a team of players (2-3 players per team) will take the role of a supply chain coordinator and is responsible in managing their goods, supplies and facilities to fulfil the demand from their customers (or the next level in the supply chain). The learning points for the basic scenario are bullwhip effect, lead time, and stock management. The game can be played by 4 to 10 players for approximately 90-120 minutes.

3.3.2. Humanitarian Logistics Scenario

The humanitarian logistics scenario adapts Multi Criteria Decision Making (MCDM) concept to determine the locations of the warehouses/distribution centers for relief supply stocks in humanitarian setting [60]. MCDM is a structured framework to analyze decision problems that involve complex multiple objectives [58, 59]. It can be used to accommodate complex decision processes involving various criteria and various decision makers with different viewpoints.

For this scenario, players are divided into different teams. Each team represents an expert team of regulators in humanitarian disaster relief. Their aim is to provide a continuous delivery of relief goods to the disaster area. To do so, each team first need to determine the optimal location for their warehouse and later to optimize their operation (delivery). Failure to deliver the required relief goods at the right time would be penalized with a failure token for the team. The team with the least failure token at the end of the game is the winner. The game can be played by 6-10 players with 60-90 min duration of play.

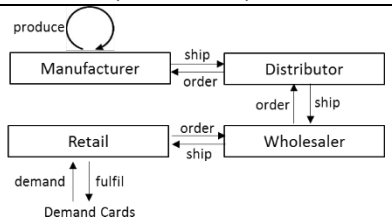
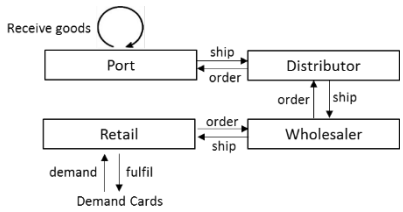
The game is divided into two stages, namely: preparation and respond stage. In the preparation stage, the players need to decide the optimal warehouse location(s) by MCDM methodology by considering a list of decision criteria, described in [60]. While in the respond stage, the players need to deliver the goods to the main disaster area. The game play at this stage is similar to the basic scenario.

3.3.3. Urban Logistics Scenario

Urban logistics scenario focuses on last mile deliveries in urban areas. It has four different roles, namely: port, distributor, wholesaler and retailer. Ideally, the game should be played by 8 players where 1 player as the port, 2 players as distributors, 2 players as wholesalers and 3 players as retailers. Each player is competing with each other to optimize their logistics activities.

This scenario has two stages, namely: preparation and respond stage. Similar to humanitarian logistics scenario, in the preparation stage, the players (except for the port) need to decide the optimal warehouse location(s) using MCDM methodology. But, the criteria used are different from the humanitarian logistics scenario. On the respond stage, the players need to deliver the goods to their customers. It is similar to the basic scenario, with the following differences: 1) the demand locations are random for each round and scattered all over the map; 2) each

Table 3. ThinkLog Scenarios

	ThinkLog		
	Basic scenario	Humanitarian logistics scenario	Urban logistics scenario
Learning objective	• To provide an overview of information flow and coordination challenges in supply chain and logistics • To introduce the bullwhip effect	• To introduce the concept of Multi Criteria Decision Making (MCDM) for selecting warehouse location for relief supply stocks in humanitarian logistics	• To introduce MCDM for selecting warehouses/collection and delivery points (CDP) and last-mile challenges in urban distribution
Game components			
Main board	Basic main board	Basic main board	Flexible main board
Demand card	Demand card shows the demand value	Demand card shows demand value	Demand card shows demand value and demand location. Note: demand locations are scattered and dispersed all over the map.
Gameplay	1 stage	2 stages, namely: • preparation stage (to determine the warehouse location) • respond stage (to deliver certain amount of relief goods to the main disaster area)	2 stages, namely: • preparation stage (to determine the warehouse location) • respond stage (to deliver certain amount of goods to the urban area)
Roles	4 roles, namely: manufacturer, distributor, wholesaler, and retailer	1 roles: regulator in humanitarian disaster relief	4 roles, namely: port, distributor, wholesaler, and retailer
Relation between roles		-	
Game setup			
Number of players	4-10 players	6-10 players	4-8 players
Game duration	90-120 minutes/session	60-90 minutes	60-90 minutes
Performance measurement	Remaining money and backlogs	Remaining money and saved lives (represented by failure token)	Remaining money and backlogs
Twists	Common disruptions such as traffic congestions and flooding may be introduced on certain part of the game		

retailer would receive the same demand value and location; 3) retailers would be able to order goods from any wholesaler; 4) wholesalers would be able to order goods from any distributor; 5) port has two dedicated warehouse locations and the delivery will start from these two locations; and 6) the players are able to consolidate the shipments from two or more orders.

3.4. Companion Digital Application

From our first interactive session, we noted some improvements were needed to maximize the players' experience, such as in auto-tracking and graphical presentation of players' actions. These improvements may help to disseminate certain concepts better. For example, by producing a graph at the end of the game session, it may help to explain the bullwhip effect more effectively and help the players to review their decisions throughout the game and how it had affected the whole supply chain performance. In addition to that, several scenarios, such as humanitarian and urban logistics scenarios, may require some complex steps and calculations to determine the best location for the warehouse. It can be very time-consuming and prone to human error if it is to be done manually.

Based on the above improvement needs, we develop a digital companion application. It is designed not only to assist in complex calculation and display information, but also to guide the players during the game, support the game master (especially the non-experience game master) and record the player's actions and decisions during the game. We believe that with the companion app, the gameplay can be more interactive and lively,

not only between players but also between the game and the players.

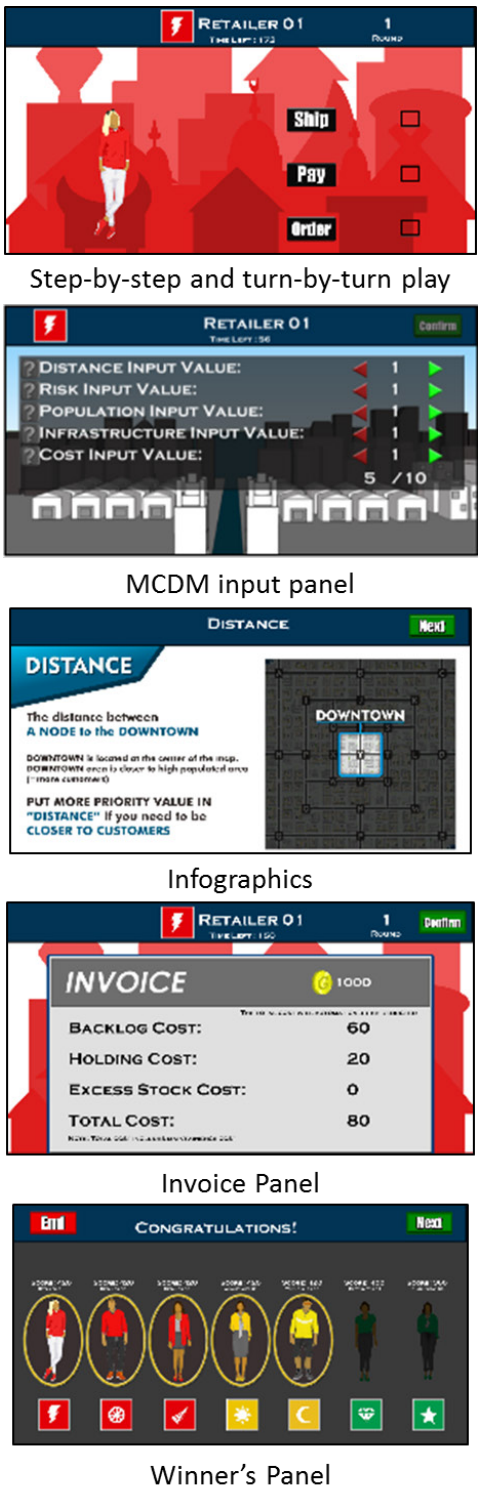


Figure 4. Examples of the Companion App Interfaces

The companion app is developed as desktop and mobile application. Several interfaces of the companion app are shown in Figure 4. Ideally, the companion app should be provided to all the players, downloadable for their digital devices. The features of the companion app include:

- Provides step-by-step, turn-by-turn play instructions to the players
- Serves as timekeeper for each player. Failing to keep to the time, the app will assume that no action is required from the player and ends the player's turn
- Captures the transaction inputs from the players and displays them as graphs at the end of the game
- Auto-generates a list of the most suitable locations for the player to choose based on the players' input for the MCDM (for scenarios that use MCDM)
- Provides a function to randomize the congestion locations for each round
- Provides infographics on the different criteria available for the MCDM
- Auto-deducts the cost incurred by each player and presenting its breakdown cost at each turn
- Auto-generates the winner(s) at the end of the game

4. Evaluation Method

To evaluate the effectiveness of ThinkLog, we conducted four interactive sessions with senior government officials and supply chain specialists as part of a three-day workshop in SCM. The sessions were held in different cities in Indonesia from August 2016 to April 2018 with different players for each session. Different sessions were played in different ThinkLog scenarios with or without the companion app. Summary of the sessions played is presented in Table 4.

In each session, we divided the players into small groups of eight to ten players where each group would have a mixed composition of players with different background. During the game session, the players were able to interact and discuss with the game masters and other players. Examples of that interaction without and with companion app are shown in Figure 5 and Figure 6 respectively.

We consider Session 1 as a trial session to measure the players' acceptance. For this session, we focused on observing the players' interaction and collecting their sentiments toward the game. In the subsequent sessions, we gathered more feedbacks from the players as well as from the game masters. These include players' sentiment, game experience, learning objective and feedback for companion app. Since the interactive sessions were run under different experimental conditions (i.e. using different scenario, different location and different game masters), we decided to treat the feedback from each session separately.

4.1. Sentiment Analysis

To gather the players' sentiments toward the game, we asked the players to provide their comments in an open-ended question. We then applied sentiment analysis method in R to evaluate these comments. Sentiment analysis has been widely used to study opinions, sentiments and emotions expressed in texts to identify positive or negative opinions based on a set of positive and negative lexicon [61].

Table 4. Summary of the Interactive Sessions

Interactive Session	Period	Location	Number of players	Using companion app	The scenario played	Feedback gathered	
						Players [feedback(#valid responses)]	Game Masters
1	August 2016	Jakarta	45	No	Basic	• Players' sentiment (13)	-
2	May 2017	Bandung	63	No	Humanitarian	• Players' sentiment (15) • Game experience (34) • Learning objective evaluation (34)	-
3	November 2017	Makassar	54	Yes	E-Commerce	• Players' sentiment (13) • Game experience (28) • Learning objective evaluation (28) • Companion app (28)	companion app
4	May 2018	Yogyakarta	44	Yes	E-Commerce	• Players' sentiment (20) • Game experience (25) • Learning objective evaluation (33) • Companion app (33)	companion app



Figure 5. Interactions between the players without companion app

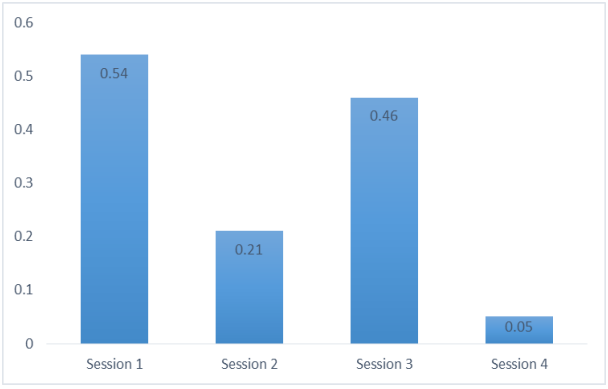
Due to limited positive and negative lexicon database in Bahasa Indonesia, for this study, we used English positive and negative lexicon database [62, 63]. Before applying the sentiment analysis method, we translated the comments/feedback in English. The survey responses and sentiment analysis score are shown in Figure 7. The result shows that the average sentiment score is above 0. This indicates positive sentiments toward the game.

4.2. Game Experience

We further evaluated the players' game experience in session 2 to 4. We developed a questionnaire comprising 9 questions focuses on four experience categories, namely: positive, negative, flow, and learning experience. Flow experience is defined as a total engagement in the game that motives the players to win the game by achieving new skills and understanding new concepts voluntarily [64], while learning experience is defined as an ability of the game to transfer its learning objectives to the players [65]. The questionnaire used the 4-point Likert scale (4=strongly agrees, 1=strongly disagrees). Summary of the responses is shown in Figure 8.



Figure 6. Interactions between the players with companion app



* = score <0 represents negative opinion, score =0 represents neutral opinion and score >0 represents positive opinion

Figure 7. The Players' Sentiment Analysis Score

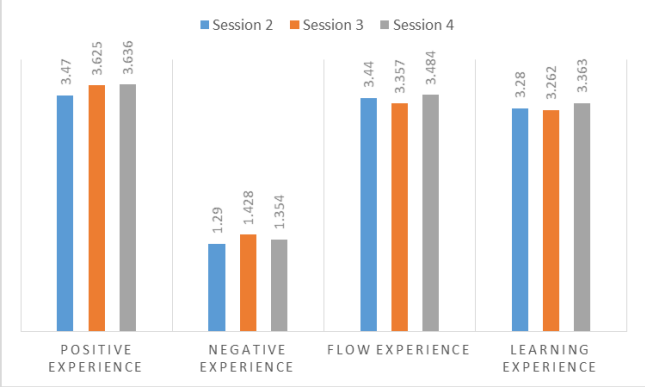


Figure 8. The Players' Overall Game Experience

Figure 8 shows that the average score for two experience categories (positive experience and flow experience) is above 3.3 and the average score for negative experience is below 1.5 for all the game sessions. It indicates that the players have a good experience with the game. For the learning experience, the result shows that the average score is above 3.2 which suggests that the players are able to understand the learning objectives.

4.3. Learning Objective

The learning objective is evaluated in the sessions 2 to 4. Session 2 used the humanitarian logistics scenario, while session 3 and 4 used the urban logistics scenario. The intended learning objectives for these two scenarios are described in Table 3. These intended learning objectives are to be evaluated against the learning points from the players' perspective.

In session 2, we collected the players' feedback by asking each group of players to list down learning points from the game after they have completed the game. A sample of those lists is shown in Figure 9. In this figure, the players mentioned that their learning points from the game were understanding criteria for determining the location of warehouse/distribution center for relief supply stocks in humanitarian logistics. Those criteria are distance, disaster location, congestions, infrastructure and maintenance cost.

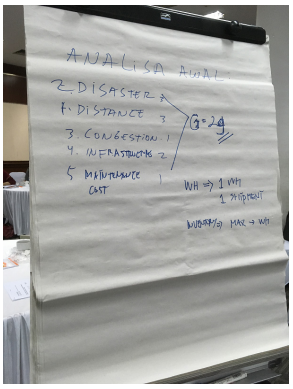


Figure 9. Sample of poster for the players' learning points

The summary of the learning points from the players' perspective in session 2 is as follows:

1. There are a lot of criteria to consider when determining the location of warehouse/distribution center, such as distance to the demand area, congestion, access to infrastructure and cost.
2. The decision makers need to consider some trade-off between those criteria.
3. Communication and coordination are essential to fulfil the demand effectively and efficiently.

These learning points are aligned with the intended learning objectives for humanitarian scenario. In session 3 and 4, we gathered the players' learning points individually. Each player was able to list down one or more learning points. Since we were playing the same scenario in these two sessions, we combined the players' learning points and summarized it in Table 5. From the result, we learnt that the players' learning points are aligned with the intended learning objectives for urban logistics.

4.4. Feedback on Companion App

The companion app was first introduced at interactive session 3 and 4. We gathered feedbacks on the companion app from the players and game masters. For the players, we asked them to evaluate the companion app usefulness in two criteria, namely: helping the players to understand the rule of the game and helping the players to learn about SCM concepts and used a 4-point Likert scale (4=strongly agrees, 1=strongly disagrees) to evaluate the criteria. We received 61 valid responses from interactive session 3 and 4. The average score for both criteria is above 3 (3.1 for helping the players to understand the rule and 3.2 for helping the players to learn). It indicates the companion app was well received by the players and it helps with the understanding and learning of the supply chain concepts embedded in the game.

Table 5. Learning Points absorb by the Players in the Third and Fourth Sessions

Learning Points	Number of Players
Importance of warehouse location	24
Warehouse selection criteria	11
Order planning and fulfillment	10
Stakeholders' coordination	8
Inventory management	6
Supply chain strategy and management	6
Warehouse and route optimization	5
Cost management	5
Demand forecasting	3
Communication	3
Simulation	1
Supply chain risk	1

To gather feedbacks from the game masters, we conducted group discussions after the game sessions were completed. From the discussions, we learnt that the companion app is able to speed up and smoothen the game process, but the interface may need to be improved especially for small screen devices. The information shown in the companion app may also need to be filtered or re-arranged to avoid information overload that would confuse the players.

5. Incorporating ThinkLog into Supply Chain Courses

With its expandable capability, ThinkLog has great potential to be incorporated into supply chain courses with large number of students. ThinkLog scenario can be adjusted to fit in different

courses. For example, the basic scenario is suitable to provide an introduction of SCM, while the humanitarian logistics scenario is suitable for more advance courses or special topics in SCM.

However, there are several challenges to incorporate ThinkLog in courses, especially for classes with a large number of students. Firstly, board game is usually played in small groups to encourage interaction within the group itself [66]. With limited number of game resources (i.e. boards and game masters), it may be impossible to divide large number of students into smaller groups. Secondly, it would be difficult to maintain the level of experience from all the students because the lecturer may not be able to engage with all of the students. Hence, it would be difficult for the students to relate their game experience with the course material.

To mitigate these challenges, we adjust the setup and components of ThinkLog as follows.

1. **Group formation**. The students are divided in two competing groups. Each group would have its own board. Within each group, students are equally divided to take a particular role (i.e. manufacturer, distributor, wholesaler, and retailer) in the game.
2. **Main board.** Instead of using real board game, a digital board, projected on the big screen is used to improve visibility for all the students.
3. **Demand card.** Demand card will be randomly generated using the companion app and shown on the big screen.
4. **Game Master**. The course lecturer would replace the role of the game master.
5. **Companion app**. The app would help to keep track of the goods' movement and show it on the digital screen. Virtual money is also used to replace the "game" money.

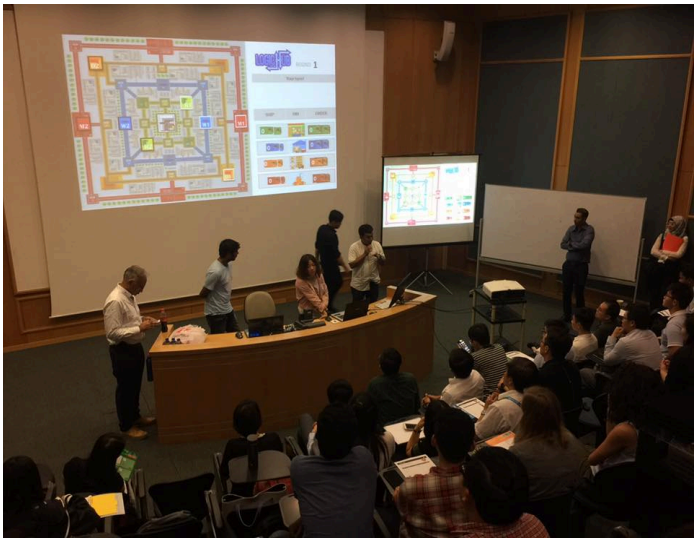


Figure 10. ThinkLog session in a supply chain course in National University of Singapore

We applied these adjustments (with the exception of the Companion app) and tested it on one supply chain course for master students in the National University of Singapore. The class was conducted on August 2017 and attended by 90 students from various backgrounds. Most of the students had background knowledge about SCM. During the game session, we observed that the students were able to discuss within their group and able to raise questions to the lecturer. An example of that interaction is shown in Figure 10.

Based on our observation and discussion with some of the students after the game session, we conclude that the students enjoy the game and highly accept it as part of their course. It's also apparent that the game helps them to understand about the course content better by applying it in the simulated environment within the game.

6. Summary and Future Work

In this paper, we focus our study on developing and evaluating an expandable board game, named ThinkLog, as an interactive tool to facilitate learning in Supply Chain Management (SCM). There are four main components in ThinkLog, namely: main board, demand card, gameplay and game master. One unique feature in ThinkLog is that this game can be extended to generate different scenarios for various concepts in SCM without changing the basic game structure. The gameplay can be adjusted to include various scenarios for different learning objectives. To date, we have extended the ThinkLog basic scenario to include two other SCM concepts, namely: humanitarian and urban logistics.

Four interactive sessions as part of three-day SCM workshops were conducted to evaluate ThinkLog. We use different scenarios for each session. Based on our observation, survey and sentiment analysis of the interactive sessions, we learnt that ThinkLog is highly accepted by the players to deepen their understanding of SCM concepts. The result of the learning objective evaluation in sessions 2, 3 and 4 also confirms that the players were able to absorb the learning objectives of the game regardless of the scenario played in the sessions. We also tested our proposed game in a big class environment as part of a supply chain course at the National University of Singapore. The acceptance and learning objective evaluation of the players resonates throughout the game session.

Nonetheless, we see four possible extensions that we would like to study in the near future. First, we would like to test our game in wider audiences and incorporate the game in different supply chain management courses. Further adjustment to the game setup may be needed to cater for different class environment. Second, we would like to embed more scenarios into the game such as healthcare logistics. We may also want to introduce supply chain credibility concept where each role or player would have its own rating or random rating to describe actual distribution processes. It would enrich the game as a platform for learning SCM concepts. Third, we would like to develop a more comprehensive evaluation method to measure the learning outcome of the game. It may include having a third party team to objectively evaluate the players' learning points and compare it with the intended learning objectives of the game. Last but not least, we would like to conduct comprehensive data analysis for each game session to further understand the players' behavior and their actions. For example, the data captured from companion app would help to indicate whether the time allocated for the game session is sufficient or if there is a correlation between the players' background knowledge and the decisions made.

Conflict of Interest

The authors declare no conflict of interest related to this paper.

Acknowledgment

The work is part of solutions/case studies developed as part of "Temasek Foundation International – National University of Singapore Urban Transportation Management Programme in Indonesia" programme. It is supported by Temasek Foundation International and Coordinating Ministry for Economic Affairs of Indonesia.

References

[1] Lindawati, E. Nugroho, R. Fredericco, Z. Abdul Rahim and R. de Souza, "ThinkLog: Interactive Learning for Supply Chain Management," in *IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Hong Kong, 2017.

[2] S. Tobias, J. Fletcher and A. Wind, "Game-based learning," in *Handbook of research on educational communications and technology*, New York, Springer, 2014, pp. 485-503.

[3] M. Ma, A. Oikonomou and L. Jain, "Innovations in Serious Games for Future Learning," in *Serious Games and Edutainment Applications*, London, Springer, 2011, pp. 3-7.

[4] T. Susi, M. Johannesson and P. Backlund, "Serious games: An overview," Technical Report, School of Humanities and Informatics, University of Skövde, Skövde, Sweden, 2007.

[5] S. de Freitas and F. Liarokapis, "Serious Games: A New Paradigm for Education?," in *Serious Games and Edutainment Applications*, London, Springer, 2011, pp. 9-21.

[6] J. Riedel and J. Hauge, "State of the art of serious games for business and industry," in *Concurrent Enterprising (ICE), 2011 17th International Conference* , 2011.

[7] J. Breuer and G. Bente, "Why so serious? On the Relation of Serious Games and Learning," *Journal for Computer Game Culture,* vol. 4, no. 1, pp. 7-24, 2010.

[8] M. Garcia-Ruiz, J. Tashiro, B. Kapralos and M. Martin, "Crouching Tangents, Hidden Danger: Assessing Development of Dangerous Misconceptions within Serious Games for Healthcare Education," in *Gaming and Simulations: Concepts, Methodologies, Tools and Applications*, Hershey, PA, Information Resources Management Association, 2011, pp. 1712-1749.

[9] S. de Freitas and S. Jarvis, "Towards a development approach for serious games," in *Games-based learning advancements for multi-sensory human-computer interfaces: Techniques and effective practices*, T. Connolly, M. Stansfield and E. Boyle, Eds., Hershey, PA, IGI Global, 2008, pp. 215-231.

[10] M. Graafland, J. Schraagen and M. Schijven, "Systematic review of serious games for medical education and surgical skills training," *British Journal of Surgery,* vol. 99, no. 10, pp. 1322-1330, 2012.

[11] C. Lim and H. Jung, "A study on the military Serious Game," *Advanced Science and Technology Letters,* vol. 39, pp. 73-77, 2013.

[12] A. Gómez-Rodríguez, J. González-Moreno, D. Ramos-Valcárcel and L. Vázquez-López, "Modeling serious games using AOSE methodologies," in *11th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE*, 53-58, 2011.

[13] J. Sterman, "Modeling Managerial Behaviour: Misperceptions of Feedback in a Dynamic Decesion Making Experiment," *Management Science,* vol. 35, no. 3, pp. 321-339, 1989.

[14] F. a. S. R. Chen, "The stationary beer game," *Production and Operations Management,* vol. 9, no. 1, pp. 19-30, 2000.

[15] E. Anderson and D. Morrice, "A Simulation Game for Teaching Service-Oriented Supply Chain Management: Does Information Sharing Help Managers with Service Capacity Decision?," *Production and Operations Management,* vol. 9, no. 1, pp. 40-55, 2000.

[16] S. van Houten, A. Verbraeck, S. Boyson and T. Corsi, "Training for today's supply chains: an introduction to the distributor game," in *The 37th Winter Simulation Conference*, 2005.

[17] T. Corsi, S. Boyson, A. Verbraeck, S. Van Houten, C. Han and J. Macdonald, "The real-time global supply chain game: New educational tool for developing supply chain management professionals," *Transportation Journal,* vol. 45, no. 3, pp. 61-73, 2006.

[18] L. Rieber, "Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games," *Educational Technology Research and Development,* vol. 44, no. 2, pp. 43-58, 1996.

[19] J. Knight, S. Carly, B. Tregunna, S. Jarvis, R. Smithies, S. de Freitas, K. Mackway-Jones and I. Dunwell, "Serious gaming technology in major incident triage training: A pragmatic controlled trial," *Resuscitation Journal,* vol. 81, no. 9, pp. 1174-1179, 2010.

[20] H. Hou, "Integrating cluster and sequential analysis to explore learners' flow and behavioral patterns in a simulation game with situated-learning context for science courses: A video-based process exploration," *Computers in human behavior,* vol. 48, pp. 424-435, 2015.

[21] J. Hamari, D. Shernoff, E. Rowe, B. Coller, J. Asbell-Clarke and T. Edwards, "Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning," *Computers in Human Behavior,* vol. 54, pp. 170-179, 2016.

[22] L. Zhou, Y. Xie, N. Wild and C. Hunt, "Learning and practising supply chain management strategies from a business simulation game: a comprehensive supply chain simulation," in *Simulation Conference, 2008*, 2008.

[23] K. Katsaliaki, N. Mustafee and S. Kumar, "A game-based approach towards facilitating decision making for perishable products: An example of blood supply chain," *Expert Systems with Applications,* vol. 41, no. 9, pp. 4043-4059, 2014.

[24] T. Nilsen, S. Linton and J. Looser, "Motivations for augmented reality gaming," in *Game Developers Conference, FUSE 4*, New Zealand, 2004.

[25] S. Gobel, "Serious Game Application Examples," in *Serious Game: Foundations, Concepts, and Practice*, R. Dörner, S. Göbel, W. Effelsberg and J. Wiemeyer, Eds., Darmstadt, Wiesbaden, Springer, 2016, pp. 319-405.

[26] R. Dörner, S. Göbel, W. Effelsberg and J. Wiemeyer, Serious games: foundations, concepts and practice, Darmstadt, Wiesbaden: Springer, 2016.

[27] E. Nugroho and A. Hariyadi, "Creative Learning and Research Through Modern Board Game," in *Triple Helix 10 th International Conference*, Bandung, 2012.

[28] B. Lamey and C. Bristow, "The Game as a Talking Tool: Using a Board Game for Creative Consultation, Engagement and Inclusion," in *Creative Education, Teaching and Learning: Creativity, Engagement and the Student Experience*, G. Brewer and R. Hogarth, Eds., London, UK, Palgrave Macmillan, 2015, pp. 243-.

[29] J. Zagal, J. Rick and I. Hsi, "Collaborative games: Lessons learned from board games," *Simulation & Gaming,* vol. 37, no. 1, pp. 24-40, 2006.

[30] A. d'Astous and K. Gagnon, "An inquiry into the factors that impact on consumer appreciation of a board game.," *Journal of Consumer Marketing,* vol. 24, no. 2, pp. 80-89, 2007.

[31] M. Grechus and J. Brown, "Comparison of individualized computer game reinforcement versus peer-interactive board game reinforcement on retention of nutrition label knowledge.," *Journal of Health Education,* vol. 31, no. 3, pp. 138-142, 2000.

[32] J. Lennon and D. Coombs, "The utility of a board game for dengue haemorrhagic fever health education," *Health Education,* vol. 107, no. 3, pp. 290-306, 2007.

[33] S. Amaro, A. Viggiano, A. Di Costanzo, I. Madeo, A. Viggiano, M. Baccari, E. Marchitelli, M. Raia, E. Viggiano, S. Deepak and M. Monda, "Kaledo, a new educational board-game, gives nutritional rudiments and encourages healthy eating in children: a pilot cluster randomized trial," *European journal of pediatrics,* vol. 165, no. 9, pp. 630-635, 2006.

[34] W. Bartfay and E. Bartfay, "Promoting health in schools through a board game," *Western Journal of Nursing Research,* vol. 16, no. 4, pp. 438-446, 1994.

[35] A. Vahed, "The tooth morphology board game: an innovative strategy in tutoring dental technology learners in combating rote learning," in *2nd European Conference on Games Based Learning*, 2008.

[36] P. Ogershok and S. Cottrell, "The pediatric board game," *Medical Teacher,* vol. 26, no. 6, pp. 514-517, 2004.

[37] S. Fukuchi, L. Offutt, J. Sacks and B. Mann, "Teaching a multidisciplinary approach to cancer treatment during surgical clerkship via an interactive board game," *The American journal of surgery,* vol. 179, no. 4, pp. 337-340, 2000.

[38] I. Mayer, L. Carton, M. de Jong, M. Leijten and E. Dammers, "Gaming the future of an urban network," *Futures,* vol. 36, no. 3, pp. 311-333, 2004.

[39] E. Speelman, L. García-Barrios, J. Groot and P. Tittonell, "Gaming for smallholder participation in the design of more sustainable agricultural landscapes," *Agricultural Systems,* vol. 126, pp. 62-75, 2014.

[40] K. Slegers, S. Ruelens, J. Vissers and P. Duysburgh, "Using game principles in ux research: A board game for eliciting future user needs," in *33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[41] A. Van Ackere, E. Larsen and J. Morecroft, "Systems thinking and business process redesign: an application to the beer game," *European Management Journal,* vol. 11, no. 4, pp. 412-423, 1993.

[42] F. Jacobs, "Playing the beer distribution game over the internet," *Production and Operations Management,* vol. 9, no. 1, pp. 31-39, 2000.

[43] J. Wisner, K. Tan and G. Leong, Principles of supply chain management: A balanced approach, Cengage Learning, 2014.

[44] N. Mustafee and K. Katsaliaki, "The blood supply game," in *Winter Simulation Conference 2010*, 2010.

[45] D. Simchi-Levi, E. Simchi-Levi and P. Kaminsky, Designing and managing the supply chain: Concepts, strategies, and cases, New York: McGraw-Hill, 1999.

[46] P. Kaminsky and D. Simchi-Levi, "A new computerized beer game: A tool for teaching the value of integrated supply chain management," *Global supply chain and technology management,* vol. 1, no. 1, pp. 216-225, 1998.

[47] A. Abdul Jabbar and P. Felicia, "Gameplay engagement and learning in game-based learning: A systematic review," *Review of Educational Research,* vol. 85, no. 4, pp. 740-779, 2015.

[48] M. Holweg and J. Bicheno, "Supply chain simulation - a tool for education, enhancement and endeavour," *International Journal of Production Economics,* vol. 78, no. 2, pp. 163-175, 2002.

[49] T. Muller, R. Müller, K. Zedel, G. Zomer and M. Engler, "Enhancing awareness on the benefits of supply chain visibility through serious gaming," in *Conference on e-Business, e-Services and e-Society*, 2015.

[50] D. R. Cruickshank and R. Telfer, "Classroom games and simulations," *Theory Into Practice,* vol. 19, pp. 75-80, 1980.

[51] M. R. Lepper and D. L. Cordova, "A desire to be taught: Instructional consequences of intrinsic motivation," *Motivation and Emotion,* vol. 16, pp. 187-208 , 1992.

[52] T. W. Malone and M. R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for learning," in *Aptitude, learning and instruction*, vol. 3, R. Snow and M. Farr, Eds., Hillsdale, NJ, Erlbaum, 1987, pp. 223-253.

[53] R. Rouse, Game design: theory & practice, Plano, TX: Wordware Publishing, Inc , 2005.

[54] M. J. Hannafin and K. Peck, The design, development and evaluation of instructional software, New York: MacMillan Publishing Company, 1988.

[55] A. Wilson, C. Broadbent, B. McGrath and J. Prescott, "Factors Associated with Player Satisfaction and Educational Value of Serious Games," in *Serious Games and Edutainment Applications*, Springer, 2017, pp. 513-535.

[56] J. Van Staalduinen and S. de Freitas, "A game-based learning framework: Linking game design and learning," in *Learning to play: exploring the future of education with video games*, M. Khine, Ed., New York, Peter Lang, 2011, pp. 29-54.

[57] S. M. Alessi and S. R. Trollip, Multimedia for learning: Methods and development, Boston: Allyn and Bacon, 2001.

[58] P. Vincke, "MCDM: Past decade and future trends—A source book of multiple criteria decision making," *European Journal of Operational Research,* vol. 26, no. 1, pp. 179-780, 1986.

[59] M. Köksalan, J. Wallenius and S. Zionts, "An Early History of Multiple Criteria Decision Making," *Journal of Multi-Criteria Decision Analysis,* vol. 20, no. 1-2, pp. 87-94, 2013.

[60] G. Timperio, G. Panchal, R. De Souza, M. Goh and A. Samvedi, "Decision making framework for emergency response preparedness: A supply chain resilience approach," in *IEEE International Conference on Management of Innovation and Technology (ICMIT)*, 2016.

[61] G. Miner, J. Elder IV and T. Hill, Practical text mining and statistical analysis for non-structured text data applications, Academic Press, 2012.

[62] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews,," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Seattle, Washington, USA, 2004.

[63] B. Liu, M. Hu and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *the 14th International World Wide Web conference (WWW-2005)*, Chiba, Japan, 2005.

[64] K. Kiili, "Evaluations of an Experiential Gaming Model," *An Interdisciplinary Journal on Humans in ICT Enviroments,* vol. 2, no. 2, pp. 187-201, 2006.

[65] H. Kelly, "How Games and Simulations Can Help Meet America's Challenges in Science Mathematics and Technology Education," in *Serious Game Design and Development: Technologies for Training and Learning*, IGI Global, 2010, pp. 117-133.

[66] S. Fukuchi, L. Offutt, J. Sacks and B. Mann, "Teaching a multidisciplinary approach to cancer treatment during surgical clerkship via an interactive board game," *The American journal of surgery,* vol. 179, no. 4, pp. 337-340, 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 105-108 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

An Electroencephalogram Analysis Method to Detect Preference Patterns Using Gray Association Degrees and Support Vector Machines

Shin-ichi Ito*, Momoyo Ito, Minoru Fukumi

Department of Science and Technology, Faculty of Science and Technology, Tokushima University, 770-8506, Japan

ARTICLE INFO

Article history:
Received: 08 August, 2018
Accepted: 14 September, 2018
Online: 22 September, 2018

Keywords:
Electroencephalogram
Preference
Favorite sounds
Simple electroencephalography
Gray association degree
Support vector machine

ABSTRACT

This paper introduces an electroencephalogram (EEG) analysis method to detect preferences for particular sounds. Our study aims to create novel brain–computer interfaces (BMIs) to control human mental (NBMICM), which are used to detect human mental conditions i.e., preferences, thinking, and consciousness, choose stimuli to control these mental conditions, and evaluate these choices. It is important to detect the preferences on stimuli. If the stimuli related to the preference can be detected, the NBCIMC can provide stimuli to the user based on their emotions by detecting their favorite stimuli. The proposed method adopted EEG recording technique, extraction techniques of EEG features and detection methods of preferences. EEG recording employs a simple electroencephalograph, for which the measurement position is the left frontal lobe (Fp1) of the brain. We assume that the differences of the EEG activities on the patterns of preference are expressed in the association between the changes of the power spectra on each frequency band of the EEG. To calculate the association, we employ the gray theory model. The EEG feature is extracted by calculating the gray association degree, then, the preferences are detect using a support vector machine (SVM). Experiments are conducted to test the effectiveness of this method, which is validated by a mean accuracy rate >88% on the favorite sound detection. These results suggest that the detection of subject's favorite sounds becomes easy when the EEG signals are analyzed while the gray associate degrees are used as the EEG feature and the SVM is used as the classifier.

1. Introduction

This paper is an extension of a paper presented at the 2018 International Conference on Electronic, Information, and Communication (ICEIC), 2018 [1]. This paper discusses the detection of human mental state conditions, proposes a method for detecting patterns of human preferences, and contains additional discussions on comparative methods.

Recently, electroencephalogram (EEG)-based interfaces (brain–computer interfaces (BCI) and/or BMI) have gained considerable research interest in many fields. This study aims to create new BMIs to control human mental (NBMICM), which are used to detect human mental conditions e.g., preferences, thinking, consciousness, choose stimuli for controlling these mental conditions, and evaluate these choices. It is important to detect the preferences on stimuli. If the stimuli related to the preference can be detected, the NBCIMC can provide stimuli to the user based on their emotions by detecting their favorite stimuli.

There are any approaches to analyze and detect the human preferences in previous studies. Sawata et. al. proposed the method to detect the individual favorite music by calculating the audio features of EEG signals during listening to the music [2]. In [3], the author analyzed the differences of the relationship between the ethnic groups and music preferences by detecting the preference on music using EEG signals and their analyses. In [4], the authors analyzed the relationship between the haptic preference and the Gamma EEG to detect the features of the EEG signals. However, these previous studies employed the electroencephalograph that had many electrodes to record the EEG and it may be difficult to use the interface using the EEG on daily basis [5]. This paper used a simple device that has a single dry-type electrode. The sensing position is on the left lobe. The prefrontal cortex activities are changes when a human mental state is changed [6,7], therefore, EEG activities in the prefrontal pole are variable. Furthermore, the

*Shin-ichi Ito, 2-1 Minami-josanjima Tokushima Japan, +81886569858 & Email: s.ito@tokushima-u.ac.jp

EEGs of the frontal cortex activities vary from person to person [8,9]. We propose a method using single-point sensing to analyze EEGs and mitigate the adverse effects of such individual differences.

To analyze EEG activities, there are numerous approaches [10]. In [3], [4], the authors calculated the power spectrum of EEG signals to extract the features. In [11], [12], the authors employed principal component analysis for extracting EEG features [11,12]. In [13], the authors used independent component analysis to extract the features. Then, In [13,14], the researchers employed *k*-nearest neighbor (*k*NN) to detect and/or classify the EEG features. Blankertz et. al. used linear discriminant analysis (LDA) as classifier [15]. In [12], the researcher employ artificial neural networks (ANN) to classify the EEG signals. In [16], the author used self-organizing map (SOM) to recognize the human emotions. In [2], [11], [17] the authors employed support vector machine (SVM) to analyze human preference and EEG signals, respectively.

Here, we assumed that relationships among the power spectra of the frequency bands become unique when creating the preferences on stimuli and different on each preference. This paper employed the gray associate degree calculation technique [14,18] to compute the relationships. Also, the gray associate degree calculation technique can reduce the noise signals in the EEG signals because of one of statistical processing models. The SVM classifier was employed to detect the preferences because the previous studies had good results to detect the preference using the SVM. We were able to detect the subject's preference patterns by analyzing the EEG activities during listening to certain sounds. If the stimuli are sounds and the favorite sounds are detected by analyzing the EEG signals, the NBCIMC can give user healing sounds to be comfortable blanket by detecting the favorite sounds. Finally, we conducted experiments using real EEG data for testing the effectiveness of the proposed method.

2. Proposed method

The proposed method adopted EEG recording technique, extraction techniques of EEG features and detection methods of preferences, as illustrated in Figure 1.

MindTune (MT) was used to record the EEG signals. This device has a dry-type sensor and a few electrodes installed in a pair of headphones. For the EEG recordings, the left ear and Fp1 in the international 10-20 system are reference and exploring electrodes, respectively (Figure 2). The timetable of each EEG recording was 15 seconds (no listening) and 15 seconds (listening to sound) as a set, as shown in Figure 3. The recorded EEG signals were transformed into the power spectra using a fast Fourier transform. Then, the power spectra on each frequency band were based on rhythm of brain activities. This paper divided into four frequency bands; theta, low-alpha, high-alpha and beta. The range of frequency bands are 4–7Hz, 8–9Hz, 10–12Hz and 13–24Hz, respectively. After the EEG recording, the subject completed an easy questionnaire on the preference evaluation of the sounds. The preference patterns were based on the responses to the questionnaire, of which the criterion and the preference patterns (indicated in parentheses) were based on whether the subject liked

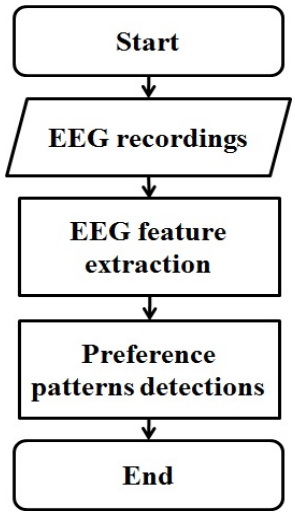


Figure 1: Procedure of the proposed method.

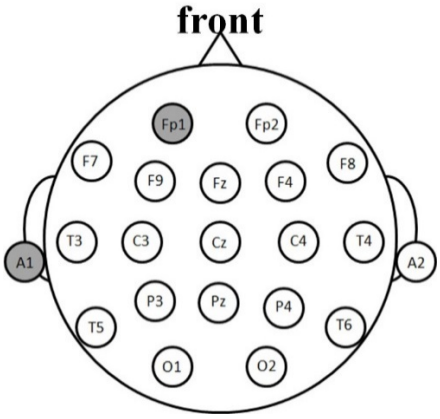


Figure 2: International 10-20 system. Reference electrode is at left earlobe (A1) and exploring electrode is at left prefrontal pole (Fp1).

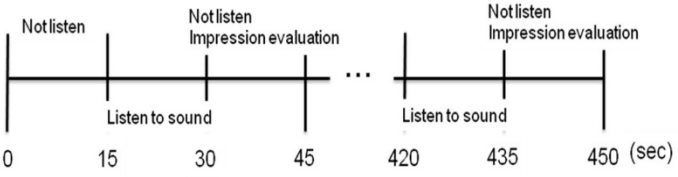


Figure 3: Timetable of each EEG recording.

(FavoriteSound), disliked (DislikeSound), or felt neutral toward, i.e., neither liked nor disliked (Other), a particular sound.

The gray associate degree was calculated to extract the EEG features based on the time variations of the power spectra on each frequency band of the EEG because we assumed that the differences of the EEG activities on the preference patterns are expressed in the association between the changes of the power spectra. The eigenvectors were calculated to extract the gray association degree from the gray relationship coefficients. We regarded the eigenvector of the greatest eigenvalue as the EEG feature. The matrices of the discrete time series data of the power spectra of the EEG frequency bands (EEG) and the gray relational coefficient (GRC) are defined as follows:

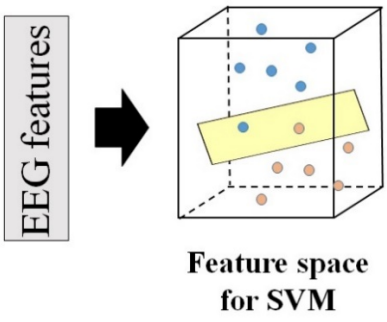


Figure 4: Classifier for detecting preference patterns using the SVM.

$$EEG = \begin{bmatrix} EEG_{\theta}(1) & EEG_{\theta}(2) & \cdots & EEG_{\theta}(T) \\ EEG_{low\alpha}(1) & EEG_{low\alpha}(2) & \cdots & EEG_{low\alpha}(T) \\ EEG_{high\alpha}(1) & EEG_{high\alpha}(2) & \cdots & EEG_{high\alpha}(T) \\ EEG_{\beta}(1) & EEG_{\beta}(2) & \cdots & EEG_{\beta}(T) \end{bmatrix} \quad (1)$$

$$GRC = \begin{bmatrix} \Gamma_{11} & \cdots & \Gamma_{14} \\ \vdots & \ddots & \vdots \\ \Gamma_{41} & \cdots & \Gamma_{44} \end{bmatrix} \quad (2)$$

subject to

$$\Gamma_{ij} = \frac{1}{T}\sum_{k=1}^{T} R_{ij}(k)$$

where EEG and T denote the power spectra and the maximum time, respectively. R is the gray relational coefficient. i and j are the frequency bands. R is defined as

$$R_{ij} = \frac{\Delta_{\min.} + \upsilon\Delta_{\max.}}{\Delta_{ij}(k) + \upsilon\Delta_{\max.}} \quad (3)$$

subject to

$$\begin{cases} \Delta_{\min.} = \min_{\forall j \in i} \min_{\forall k} \left\| EEG_i(k) - EEG_j(k) \right\|, \\ \Delta_{\max.} = \max_{\forall j \in i} \max_{\forall k} \left\| EEG_i(k) - EEG_j(k) \right\| \end{cases}$$

where υ and Δ_{ij} indicate the distinguishing coefficient ($\upsilon \in [0,1]$) and $\left\| EEG_i(k) - EEG_j(k) \right\|$, respectively. $EEG_i(k)$ and $EEG_j(k)$ denote the reference and comparison matrices, respectively.

In preference detection, the kernel SVM classifier (shown in Figure 4) using the RBF kernel function was employed to detect the preference patterns because the SVM classifier was employed to detect and/or classify the EEG features in previous studies [2,11,17]. The previous studies could obtain good results using the SVM classifier. Then, this paper employed a one-versus-rest strategy for detecting the preferences. An N -fold cross-validation method was used for testing. Finally, this paper calculated the mean accuracy rate for the preference detection ($MeanAc$) to confirm the effectiveness of the proposed method:

$$MeanAc = \frac{1}{N}\sum CrrectNum / SampleNum \quad (4)$$

where $CorrectNum$ is the total number of correct answers $SampleNum$ means the total sample number.

3. Experiments

The subjects were four males (average age: 22.5 years) and one female (age: 22 years). The EEG device was positioned on the forehead of each subject. The EEG was recorded more than once in the laboratory with ongoing background noise. The subjects listened to 15 kinds of the sounds in an experiment. The sounds comprised window bells sound, helicopter noise, fire engine , grade crossing, scotch tape, cicada buzz, bush warbler buzz, mosquito, roar of waves, soda water, unwrapping the paper, fireworks, train noise, frictional noise of styrene foams and drill noise, respectively. The experiments were repeated five times for each subject, producing a total of 375 sounds that were played. Thus, T and υ for the EEG feature extraction were 15 and 0.5, respectively, while C and γ for the kernel SVM were 2^{-5} and 2^{-12}, respectively. Then, the number of N for cross validation was 5.

Table 1 shows the experimental results for the means and standard deviations of the accuracy rates of the detections of FavoriteSound (Favorite), DislikeSound (Dislike), and Other. The comparative methods, PCA, and the power spectra were used to extract the feature. The nearest neighbor (1NN) method was used as classifier to detect the preferences. The EEG features of the comparative methods were shared by the proposed method. Gray, PCA, and Freq were the gray association degrees, which were the results of the PCA and the power spectra of the EEG (Equation 1), respectively, as the EEG features. SVM and kNN were the classifiers used by SVM and 1NN, respectively, for detecting the preference patterns.

For all methods (Gray + kNN, PCA + SVM, PCA + kNN, Freq + SVM, and Freq + kNN) other than our proposed method, the mean detection accuracies of FavoriteSound were the highest with accuracies of 88.27, 76.53, 88, 77.34, 87.47 and 82.57, respectively. Using our proposed method (Gray + SVM), the mean detection accuracies of FavoriteSound, DislikeSound, and Other were the highest with accuracies of 88.27, 52.27, and 59.6, respectively, as compared with the comparative methods (Gray + kNN, PCA + SVM, PCA + kNN, Freq + SVM, and Freq + kNN). Moreover, the standard deviations of the detection accuracies of

Table 1: Means and standard deviations of the accuracy rates of detections of FavoriteSound, DislikeSound, and Other.

(%)	Favorite	Dislike	Other
Gray + SVM (Proposed method)	**88.27 ± 0.01**	**52.27 ± 0.01**	**59.6 ± 0.01**
Gray + kNN	**76.53 ± 0.01**	50.67 ± 0.02	49.64 ± 0.01
PCA + SVM	**88.0 ± 0.01**	50.67 ± 0.01	59.47 ± 0.01
PCA + kNN	**77.34 ± 0.01**	47.46 ± 0.02	50.93 ± 0.01
Freq + SVM	**87.47 ± 0.02**	48.8 ± 0.01	52.27 ± 0.01
Freq + kNN	**82.57 ± 0.03**	47.74 ± 0.02	51.2 ± 0.01

Favorite, Dislike, and Other were 0.03, 0.02, and 0.01, respectively. The mean detection accuracies of Dislike and Other were 60% or less.

4. Discussions

Using all methods with accuracies of 76.53% or more, we confirmed that the mean detection accuracies of FavoriteSound were the highest. These results suggest that the favorite stimuli are detected by analyzing the left frontal cortex activities because the activities of the left frontal cortex on favorite stimuli become unique. Bajoulvand et. al. showed that frontal cortex activities had a close relation to human preference because the sensing position was frontal lobe and folk music preference was analyzed based on EEG analysis [3]. Therefore, this result shows the validity of our consideration.

We confirmed that the mean detection accuracies of all preferences obtained by the proposed method were higher than those obtained by the comparative methods. These results suggest that the relationships among EEG frequency bands become unique when creating the preferences on stimuli and they are different on each preference. Therefore, the detection of a subject's favorite sounds becomes easy when the EEG signals are analyzed while the gray associate degrees are used as the EEG feature and the SVM is used as the classifier. Sawata et. al. showed that favorite music was detected using SVM with mean accuracies of 83.6% or more, although 12 channels were used to record the EEG signals [2]. Therefore, this result shows the validity of our consideration.

The standard deviations of the detection accuracies for all methods were 0.03 or less, which suggest that the EEG signals and the standard deviations of the detection accuracies became high when the distributions of the extracted EEG features and the distributions related to each preference pattern were stable and the left frontal cortex activities related to the preference were widely varied.

The means of the detection accuracies of Dislike and Other were 60% or less, which were substantially lower than for FavoriteSound. These results suggest that detection is difficult for negative stimuli, such as DislikeSound, and borderline cases, such as Other, when the left frontal pole is the sensing position for the EEG analysis. If responses to negative stimuli and/or borderline case are detected, the sensing positions must be changed.

5. Conclusions

This paper introduced an EEG analysis method to detect human preferences. The proposed method adopted EEG recording technique, extraction techniques of EEG features and detection methods of preferences. In EEG measurement, the sensing point was Fp1 (the left frontal lobe). The gray association degree was used to extract the EEG feature and an SVM was used to detect the preferences of humans for particular sounds. In order to show the effectiveness of the proposed method, experiments were conducted with real EEG data. Using all comparative methods with accuracies of 76.53% or more, the mean detection accuracies on FavoriteSound were confirmed to be the highest while the proposed method produced higher accuracies for FavoriteSound, DislikeSound, and Other. These results suggest that the detection of subject's favorite sounds becomes easy when the EEG signals are analyzed while the gray associate degrees are used as the EEG feature and the SVM is used as the classifier. However, we confirmed that the detection accuracies of Dislike and Other were 60% or less, which were substantially lower than the results for the detections of FavoriteSound. These results suggest that the detection of negative stimuli, such as DislikeSound, and borderline cases, such as Other, are difficult when the sensing position for the EEG analysis is the left frontal pole.

For future research, we will change the sensing positions to detect other preferences, such as Dislike and Other.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP17K12768.

References

[1] S. Ito, M. Ito, M. Fukumi, "An Electroencephalogram Analysis Method to Detect Preference Using Gray Association Degree" in Proc. of the 17th International Conference on Electronics, Information, and Communication (ICEIC) 2018, Honolulu, 2018. http://dx.doi.org/10.23919/ELINFOCOM.2018.8330622

[2] R. Sawata, T. Ogawa, M. Haseyama, "Human-centered favorite music estimation: EEG-based extraction of audio features reflecting individual preference" Proc. of 2015 IEEE International Conference on Digital Signal Processing, 818-822, 2015.

[3] A. Bajoulvand, R. Z. Marandi, M. R. Daliri, S. H. Sabzpoushan, "Analysis of folk music preference of people from different ethnic groups using kernel-based methods on EEG signals" Applied Mathematics and Computation, **307**, 62-70, 2017.

[4] W.Park, D.-H. Kim, S.-P. Kim, J.-H. Lee, L. Kim, "Gamma EEG Correlates of Haptic Preferences for a Dial Interface" IEEE Access, **6**, 22324-22331, 2018.

[5] Y. Wang, X. Gao, B. Hong, C. Jia, S. Gao, "Brain–Computer Interfaces Based on Visual Evoked Potentials" IEEE Eng Med Biol Mag., **27**(5), 64-71, 2008.

[6] R. J. Davison, "Anterior Cerebral Asymmetry and the Nature of Emotion" Brain Cog., **20**(1), 125–151, 1995.

[7] R. J. Davison, "Anterior Electrophysiological Asymmetries, Emotion, and Depression: Conceptual and Methodological Conundrums" Psychophysiology, **35**(5), 607–614, 1998.

[8] J. B. Allen, "Issues and Assumptions on the Road from Raw Signals to Metrics of Frontal EEG Asymmetry in Emotion" Biol Psychol., **67**(1–2), 183–218, 2004. http://dx.doi.org/10.1016/j.biopsycho.2004.03.007

[9] J. A. Coan, J. B. Allen, P. E. Mcknight, "A Capability Model of Individual Differences in Frontal EEG Asymmetry" Biol Psychol., **72**(2), 198–207, 2006. http://dx.doi.org/10.1016/j.biopsycho.2005.10.003

[10] F. Lotte, M. Congedo, F. Lecuyer, B. Arnaldi, "A Review of Classification Algorithms for EEG-based Brain-Computer Interfaces" J. Neural Eng., **4**(2), R1–R13, 2007. http://dx.doi.org/10.1088/1741-2560/4/2/R01

[11] H. Lee, S. Choi, "PCA+HMM+SVM for EEG Pattern Classification" Proc. 7th Int. Symp. on Signal Processing and its Application, **1**, 541–544, 2003.

[12] T. Hoya, G. Hori, H. Bakardjian, S. Nishimura, T. Suzuki, Y. Miyawaki, A. Funase, J. Cao, "Classification of Single-trial EEG Signals by a Combined Principal + Independent Component Analysis and Probabilistic Neural Network Approach, Proc. ICA2003, **197**, 197–202, 2003.

[13] J. F. Borisoff, S. G. Mason, A. Bashashati, G. E. Birch, "Brain-Computer Interface Design for Asynchronous Control Applications: Improvements to the LF-ASD Asynchronous Brain Switch" IEEE Trans. Biomed. Eng., **51**(6), 985–992, 2004. http://dx.doi.org/10.1109/TBME.2004.827078

[14] S. Ito, M. Ito, K. Sato, S. Fujisawa, M. Fukumi, "Preference Classification Method Using EEG Analysis Based on Gray Theory and Personality Analysis" OJCSIT, **4**(3), 276–280, 2014.

[15] B. Blankertz, G. Curio, K. R. Muller, "Classifying Single Trial EEG: Towards Brain–Computer Interfacing" Adv. Neural Inf. Process. Syst., NIPS 01, **14**, 157–164, 2002.

[16] R. Khorsrowabadi, C. Q. Hiok, A. Watab, K. A. Kai, "EEG-based Emotion Recognition Using Self-Organizing Map for Boundary Detection" Proc. of ICPR, 4242–4245, 2010.

[17] T. Felzer, B. Freisieben, "Analyzing EEG Signals Using the Probability Estimating Guarded Neural Classifier" IEEE Trans. Neural Syst. Rehabil. Eng., **11**(4), 361–371, 2003.

[18] K.-L. Wen, Gray Systems: Modeling and Prediction, Alibris, Amazon, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 108-115 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Learning Personalization Based on Learning Style instruments

Alzain Alzain[*,1], Steve Clark[1], Gren Ireson[2], Ali Jwaid[1]

[1]*School of Science and Technology, Nottingham Trent University, Clifton Campus NG11 8NS, Nottingham, United Kingdom*

[2]*School of Physics & Astronomy, University of Nottingham, University Park NG7 2RD, Nottingham, United Kingdom*

ARTICLE INFO

Article history:
Received: 11 April, 2018
Accepted: 29 May, 2018
Online: 07 June, 2018

Keywords :
Learning Style
Learning Style Instruments
Adaptive Education Systems

ABSTRACT

Adaptive education systems (AES) are considered one of the most interesting research topics in technology-based learning strategies. Since students have different abilities, needs and learning styles, we should fit the curriculum and teaching activities to these different learning styles. This study investigates the impact of using LAES (Libyan Adaptive Education System) on the performance of students. An ALSI (Arabic Learning Style Instrument) was integrated into the LAES system to investigate learning preferences of students. The student models are constructed according to the results obtained using this instrument (ALSI). Three experimental studies were then conducted to investigate the impact of the LAES system on the performance of students. The results reveal the students who have learnt using the LAES system were more successful than others who learnt without, in terms of the knowledge gained.

1. Introduction

This paper is an extension of work originally presented in the *2nd International Conference on Knowledge Engineering and Applications ICKEA 2017* [1]. Research on education has indicated that students have different learning preferences, abilities and needs, and learn in different ways. For example, learners with visual learning preference tend to obtain more knowledge from instructional materials that depend on visual forms of information, whereas the content will be more beneficial for the students with verbal preferences if these materials are represented using text and audio. Moreover, some learners tend to learn more through 'doing', whereas the others prefer to think and reflect. These learning preferences are often called learning styles [2].

Although it is argued that matching of teaching styles with the preferred learning styles of students will be quite useful to improve learning outcomes [2-5], it is quite clear that many researchers also believe that learners should know more details about their learning styles because this will help them to be more engaged, motivated and attracted in educational sessions [4, 6-8]. This study investigates empirically the effect of using adaptive education systems on the performance of student learning.

[*]Alzain Alzain, Nottingham Trent University,
Email: n0214235@ntu.ac.uk

This article is organised as follows: the next section discusses the related work, where the Alzain model and ALSI instrument were selected to be integrated into the proposed system in order to profile learners; the structure of the proposed adaptive system is discussed in section 3; section 4 describes the methodology of the experiments as well as the research hypothesis; the results of our research are presented in section 5 and the conclusions are discussed in the last section.

2. Related Work

2.1. Adaptive Education Systems

Although learners have different learning preferences, goals, experiences and knowledge, the traditional education systems provide the same instructional materials for all students [9]. Therefore, in considering the individual differences between students, adaptive systems have been harnessed in the education field. The educational generation of adaptive systems is called Adaptive Learning Environment (ALE) or Adaptive Educational Hypermedia System (AEHS). These systems have been defined as "technological component of joint human–machine systems that can change their behavior to meet the changing needs of their users, often without explicit instructions from their users" [10]. This generation of educational systems can provide learners with instructional materials that are adapted especially to their learning

styles, goals, experiences or the previous knowledge of the subject [9, 11, 12]. In order to know how learners prefer to learn, learning style instruments have been developed and extensively used in adaptive education systems [13]; this situation lead us to discuss the next topic, which is learning style instruments.

2.2. Learning Style Instruments

The concept of learning styles has been harnessed in most AEHS for the purpose of building up a knowledge about students and how they prefer to learn [7, 14-16]. This knowledge is usually collected throughout psychometric questionnaires called learning style instruments, and then stored in student models with the purpose of achieving the adaptation process [12]. In the past decades, a number of learning style instruments were developed to assist learners to measure their preferred learning style and to help teachers to realize the characteristics of students [13]. Many issues concerning the integration of learning style instruments into adaptive education systems have attracted the attentions of researchers from the fields of education and computer science. The following subsection explains the learning style model and instrument that is harnessed in this study.

2.3. ALSI Instrument

The Arabic Learning Style Instrument (ALSI), was developed based on the Alzain learning style model [17] to assess student preferences on (visual, verbal, passive and active) learning styles. See figure 1.

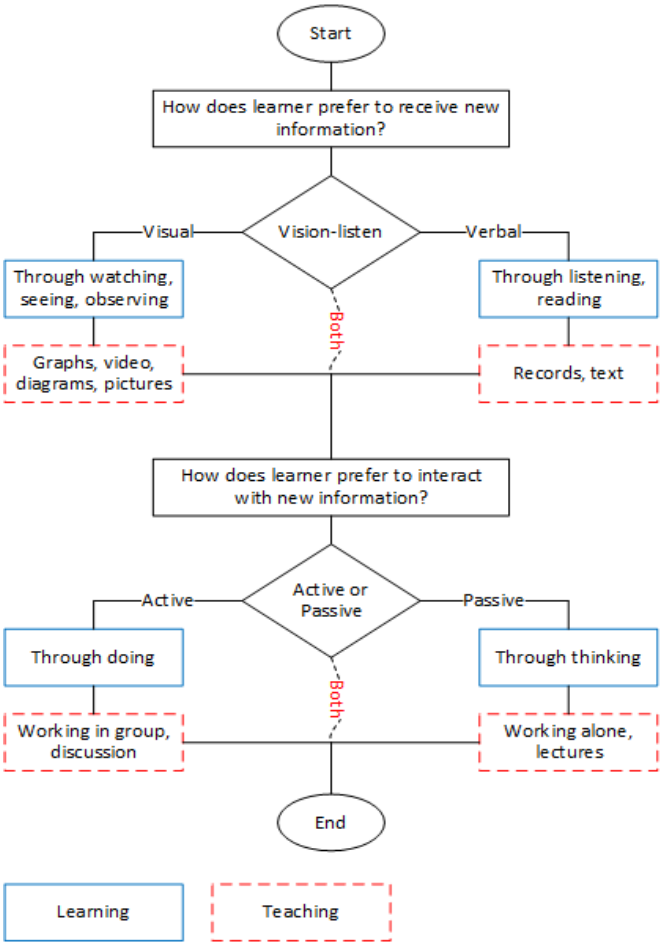


Figure 1. Alzain Learning Style Model

Based on the Alzain Learning Style Model, there are four types of combination of leaning styles. See Table 1.

Table 1. Combination of learning styles

Combination of learning styles	
1	Visual / Active
2	Visual / Passive
3	Verbal / Active
4	Verbal / Passive

These different types of combinations are considered by the use of the following elements, and the rules of each type are described below:

- Visual: get more from visual forms of information
  - More figures, graphs, charts and pictures;
  - Highlighting and colouring the important concepts;
  - Multimedia and animated demonstrations.
- Active: doing very well in groups
  - Providing discussion areas;
  - More exercises;
  - Fewer examples.
- Verbal: get more from verbal forms of information
  - Heavy textual content;
  - Audio records and files.
- Passive: thinking before doing
  - Less detailed content (summarised);
  - Giving time to think periodically;
  - More examples;
  - Fewer exercises.

The ALSI instrument consists of sixteen items, each of which has four choices, which correspond to the four learning styles. Respondents need to choose the answer(s) that best fits their preference(s) by determining the priority levels from least important (0) to most important (3), for the respective choices. The respondents are also allowed to give the same priority level for different choices at the same time.

3. LAES System

The adaptive system that employed in this study called LAES (Libyan Adaptive Education System), it is a Web-based education system. The LAES tackles the problems arising from individual differences by presenting the most suitable educational materials and activities for students. Figure 2 shows the LAES architecture, which includes four main domains:

- Content Model;
- Student Model;

- Teaching Strategies Model;
- Pedagogical Model: this model involves three components:
    - Preferences Detection Component;
    - Adaptation Component;
    - Revision Component.

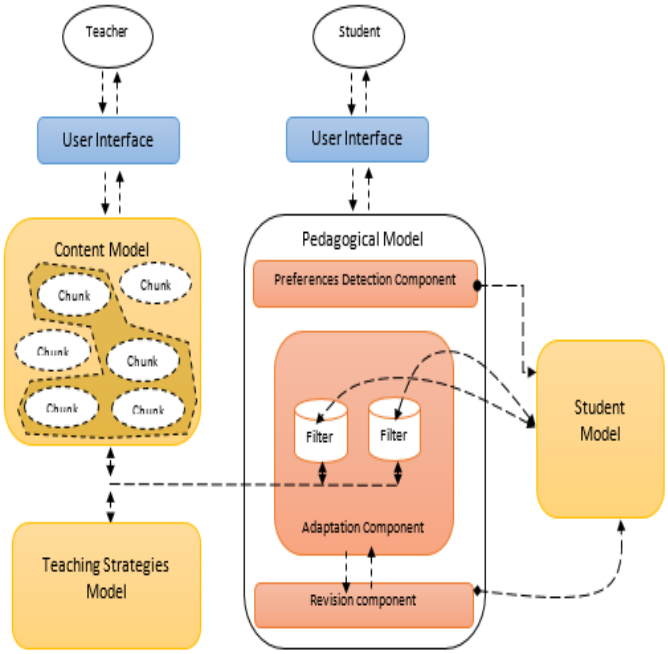


Figure 2. LAES system architecture.

3.1. Content Model

A content model includes the educational content. Typically, each course can be depicted as a tree, which consists of a set of weeks, and each week involves a number of lectures that involve a set of educational units called chunks (see Figure 3). Each unit starts with outlines and then presents the content and concludes with the summary.

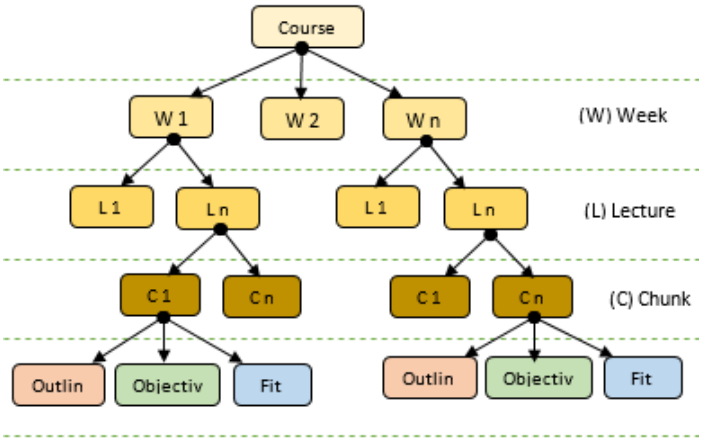


Figure 3. Content model _ LAES system

The educational materials employed in this research were designed based on the ideas of two well-known educational theories, namely Elaboration Theory and Component Display Theory (CDT) [18].

3.2. Student Model

A student model keeps the student details and their learning preferences. Accordingly, based on these details, the instructional materials and teaching strategies can be adapted to fit the learning style of the students. The student model represents a student profile, which stores all user-relevant details. These details can be divided into two main parts. While the first part summarises the learning style of students, as detected by the ALSI instrument, the second part holds the personal details of students including student name, age, email, etc. Figure 4 shows the structure of this model.

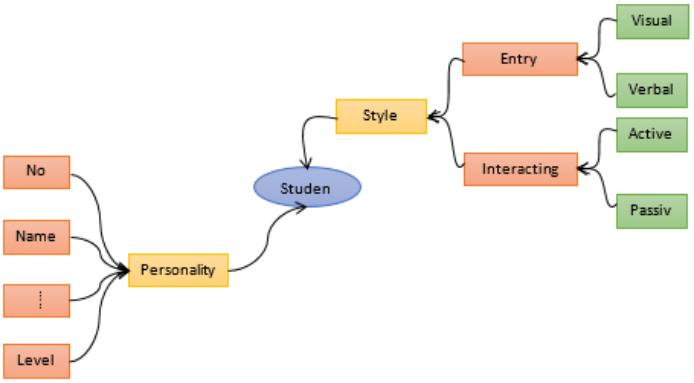


Figure 4. Student model _ LAES system

3.3. Teaching Strategies Model

This model contains a description of different teaching strategies that can be used to teach the different types of learners. Typically, each teaching strategy involves a set of activities. In this sense, the teaching strategy model can be presented as a tree (see Figure 5).

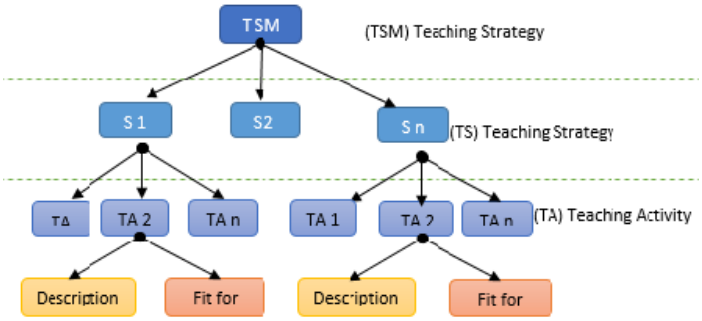


Figure 5. Teaching strategies model _ LAES system

3.4. Pedagogical Model

The pedagogical model aims to provide each individual student with the most suitable content and teaching activities. To this end, if the student is a new user, the system will direct them to fill out the learning style instrument (ALSI) to detect the student learning style, which will be stored in a student model. The learning session starts when the student is logged in. Accordingly, the LAES system recommends the most suitable content and teaching activities based on the preferred learning style of the student who is logged in.

To determine the preferred learning style of students, the procedure is as follows:

- The ALSI instrument consists of 16 questions;
- Each question has 4 choices;
- Participants need to give a priority level from 0 (least important) to 3 (most important) for each choice;
- Each choice corresponds to one preference;
- The highest score possible is 48 for each preference;
- Visual Preferences (VP) = $\sum_{Q=1}^{16} V$;
- Verbal Preferences (EP) = $\sum_{Q=1}^{16} E$;
- Active Preferences (AP) = $\sum_{Q=1}^{16} A$;
- Passive Preferences (PP) = $\sum_{Q=1}^{16} R$;
- Preferred Style of Receiving new information (PSR) = VP – EP;
    - If PSR > 0 then student has a Visual preference;
    - If PSR < 0 then student has a Verbal preference;
    - If PSR = 0 then student has equal preferences;
- Preferred Style of Interacting new information (PSI)= AP-PP;
    - If PSI > 0 then student has an Active preference;
    - If PSI < 0 then student has a Passive preference;
    - If PSI = 0 then student has equal preferences.

4. Experiment Design

The LAES system can adapt the content based on the preferred learning style of students. In order to evaluate the impact of using the LAES system on student performance, an experimental evaluation approach was used. This approach is recommended by several researchers in this field [19].

According to Alshammari [8], conducting only one experiment will not be sufficient to evaluate the adaptive system, because the number of participants and time of learning will be limited. Therefore, three different experiments were conducted, each with a different module, subject and participants. Each experiment was carried out in three sessions, and each session lasted for about 120 minutes.

In each experiment, the participants were first taught without using the LAES system, and they were asked to complete a pre-test and a post-test to know the learning outcomes. The learning outcomes were also tested in the next experimental session, in which the participants were taught using the LAES system, and the learning outcomes of two experimental sessions were compared.

4.1. Research Hypothesis

The key issue that was considered in the following three experiments was the learning outcomes, and to investigate if students who learnt using the LAES system were better off than others who learnt without the system, in terms of the knowledge gained. In the following three experiments, the following hypotheses were investigated using a paired t-test and Pearson Correlation test:

- H_1: there will be no significant difference in terms of the knowledge gained between students who learnt using the LAES system and students who learnt without it.
- H_2: there will be no significant correlation between the dimensions on learning styles.
- H_3: there will be no significant correlation between learning styles and years of computer use.

Moreover, the effect size was also tested in each experiment. The effect size is a statistical technique used with quantitative data for exploring the difference between two groups. According to Cohen [20], the effect size (Cohen's d) can lie between 0 to1, (and some, formulae yield an effect size that is larger than 1):

- From 0 to 0.20 = weak effect;
- From 0.21 to 0.50 = modest effect;
- From 0.51 to 1.00 = moderate effect;
- > 1.00 = strong effect.

5. Results

5.1. Experiment I

This experiment was conducted by the researcher with a number of undergraduate students (n = 10) studying for a (Formal Languages and Automata Theory) module.

In the experiment, the mean age of participants was 21, the minimum age was 20 and the maximum age was 23. The participants were found to be more visual and active than verbal and passive, and the majority of the participants had moderate learning preferences. Figure 6 shows the number of participants in each sub-category.

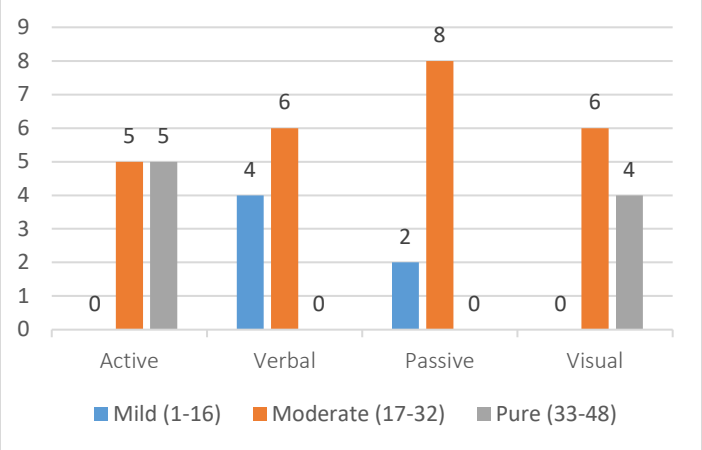


Figure 6. Participants distribution based on their learning styles _ experiment I.

In this experiment, the learning outcomes were measured. Generally, the mean participant scores when they learnt using the adaptive system (Mean = 9.60) is higher than the mean participant scores when they learnt without it (Mean = 5.30).

A dependent sample t-test was also conducted, and the results of a paired t-test showed that there was a statistically significant difference between the mean scores of the participants when they learnt using the LAES system and the mean participant scores when they learnt without it. t (10) = -2.294, p = 0.047. Therefore, it can be inferred that the students who learnt using this system were better off than others who learnt without it in terms of the knowledge gained. In this experiment, the effect size was also measured for each individual scale using Cohen's d test. The results revealed that the highest effect size (d = 1.31) was in visual style followed by the active style (d = 0.89).

Regarding the second hypothesis, which are concerned with the correlation between dimensions of learning style, the results of Pearson Correlation test showed that there was a positive significant correlation between visual and active style, r (10) = 0.71, p = 0.02. See table 2.

Table 2. Results of Pearson correlation test _ experiment I

Correlations						
		Active	Verbal	Passive	Visual	Years Of Computer Use
Active	Pearson Correlation	1	0.08	0.52	**.71***	0.12
	Sig. (2-tailed)		0.82	0.12	0.02	0.72
	N	10	10	10	10	10
Verbal	Pearson Correlation	0.08	1	-0.21	-0.23	0.44
	Sig. (2-tailed)	0.82		0.54	0.52	0.19
	N	10	10	10	10	10
Passive	Pearson Correlation	0.52	-0.21	1	0.35	-0.24
	Sig. (2-tailed)	0.12	0.54		0.31	0.50
	N	10	10	10	10	10
Visual	Pearson Correlation	**.71***	-0.23	0.35	1	-0.37
	Sig. (2-tailed)	0.02	0.52	0.31		0.28
	N	10	10	10	10	10
Years Of Computer Use	Pearson Correlation	0.12	0.44	-0.24	-0.37	1
	Sig. (2-tailed)	0.72	0.19	0.50	0.28	
	N	10	10	10	10	10
*. Correlation is significant at the 0.05 level (2-tailed).						

5.1. Experiment II

This experiment was conducted with a number of undergraduate students (n = 16) studying for a (Computer Basics) module.

In the experiment, the mean participant age was 20, the maximum age was 23 and the minimum was 18. The participants were found to be more active and visual than passive and verbal, and the majority of the participants had pure or moderate learning preferences. Figure 7 shows the number of participants in each sub-category.

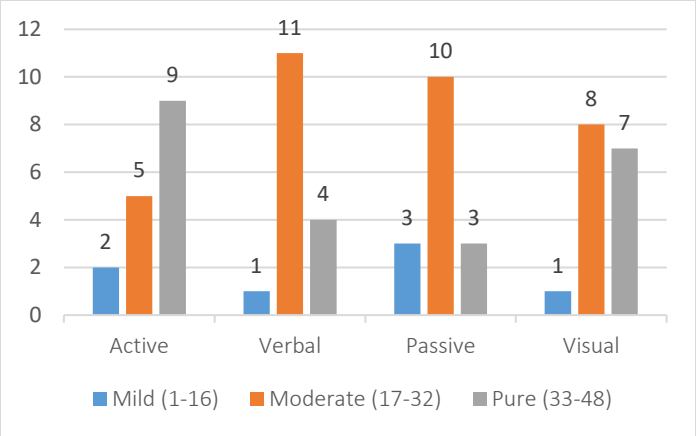


Figure 7. Participants distribution based on their learning styles _ experiment II.

In this experiment, the mean participant score when they learnt without using the LAES system (Mean = 9.94) was less than the mean participant scores when they learnt using it (Mean = 16.13). In order to investigate if there is any significant difference between the two, a dependent sample t-test was conducted. The findings showed that there was a statistically significant difference between the mean scores of the participants when they learnt using the LAES system and the mean participant scores when they learnt without it. t (16) = -2.289, p = 0.037. Therefore, it can be inferred that the students who learnt using this system were better off than others who learnt without it in terms of the knowledge gained.

The effect size was also measured for each individual scale. The results of this test revealed that the highest effect size (d = 0.77) was in the visual style followed by the verbal style (d = 0.32).

Regarding the second hypothesis, which is concerned with the correlation between dimensions of learning style, the results of Pearson Correlation test showed that there was a statistically positive significant correlation between passive and active style, r (16) = 0.79, p = 0.00, there was also positive significant correlation between passive and verbal style, r (16) = 0.76, p = 0.00. See table 3.

The results also showed that there was no significant correlation between the years of computer use and different learning styles (Table 3).

5.1. Experiment III

This experiment was conducted with a number of undergraduate students (n = 14) studying for a (Programming Languages) module.

Table 3. Results of Pearson correlation test _ experiment II

Correlations						
		Active	Verbal	Passive	Visual	Years Of Computer Use
Active	Pearson Correlation	1	0.467	**.79****	0.44	0.12
	Sig. (2-tailed)		0.06	0.00	0.08	0.64
	N	16	16	16	16	16
Verbal	Pearson Correlation	0.46	1	**.76****	-0.11	0.12
	Sig. (2-tailed)	0.06		0.00	0.67	0.65
	N	16	16	16	16	16
Passive	Pearson Correlation	**.79****	**.76****	1	0.17	0.13
	Sig. (2-tailed)	0.00	0.00		0.52	0.61
	N	16	16	16	16	16
Visual	Pearson Correlation	0.44	-0.11	0.17	1	0.12
	Sig. (2-tailed)	0.08	0.67	0.52		0.64
	N	16	16	16	16	16
Years Of Computer Use	Pearson Correlation	0.12	0.12	0.13	0.12	1
	Sig. (2-tailed)	0.64	0.65	0.61	0.64	
	N	16	16	16	16	16
**. Correlation is significant at the 0.01 level (2-tailed).						

In the experiment, the mean age was 21, the minimum age was 19 and the maximum age was 34. The participants were found to be more active and visual than passive and verbal, and the majority of the participants had pure or moderate learning preferences. Figure 8 shows the number of participants in each sub-category.

The learning outcomes were measured. Generally, the mean participant score when they learnt using the LAES system (Mean = 22.14) was higher than the mean participant scores when they learnt without using this system (Mean = 14.29).

A dependent sample t-test was also conducted, and the results of this test showed that there was a statistically significant difference between the mean score of the participant when they learnt using the LAES system and the mean participant scores when they learnt without it. t (14) = -1.724, p = 0.048. Therefore, it can be inferred that the students who learnt using this system were better off than others who learnt without it in terms of the knowledge gained.

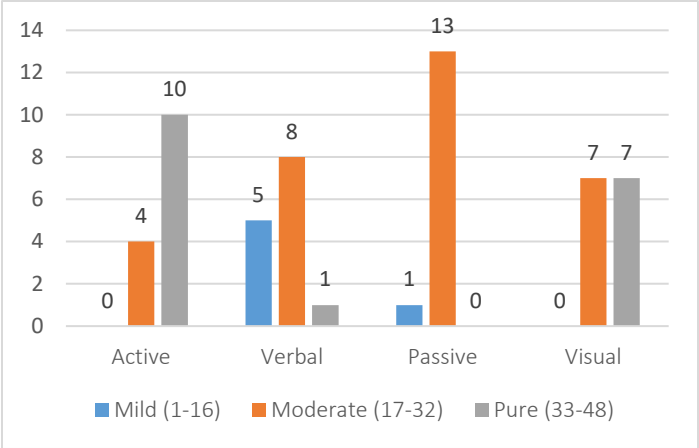


Figure 8. Participants distribution based on their learning styles _ experiment III.

The effect size was also measured for each individual scale, and the results of Cohen's d test revealed that the highest effect size (d = 0.56) was in active scale followed by the visual scale (d = 0.55).

Table 4 shows the results Pearson Correlation test, which is concerned with the correlation between dimensions of learning style, the results showed that there was a statistically positive significant correlation between visual and active style, r (10) = 0.61, p = 0.02, there was also positive significant correlation between visual style and years of computer use, r (10) = 0.58, p = 0.02.

Conclusion

This study investigated empirically the implications of using the LAES system, and the impact of that on the performance of students. It also investigated the effect size of each individual scale to identify the most affected students.

To increase the efficiency of results, three experiments were conducted. The experiments were carried out with different modules, teachers and students.

Generally, the findings indicate that using LAES system to teach students (in a matched way), based on their preferred learning style, has a positive influence on the performance of the students. The results also revealed that the visual and active students were the greatest beneficiaries from the adaptation process. A possible explanation for this result is that the existing curricula and teaching approaches are more suitable for students who are more verbal and passive than visual and active.

Regarding the first hypothesis, which is concerned with the differences in terms of the knowledge gained between students who learnt using the LAES system and students who learnt without it. In the first experiment, the results showed that the mean student scores increased from (Mean = 5.30) to (Mean = 9.60) when they learn using the LAES system. Moreover, the results of a paired t-test revealed that there was a statistically significant difference between the mean scores of the participants when they learnt using this system and the mean participant scores when they learnt

Table 4. Results of Pearson correlation test _ experiment III

Correlations						
		Active	Verbal	Passive	Visual	Years Of Computer Use
Active	Pearson Correlation	1	-0.05	0.12	**.61***	0.38
	Sig. (2-tailed)		0.86	0.67	0.02	0.18
	N	14	14	14	14	14
Verbal	Pearson Correlation	-0.05	1	0.33	-0.25	0.03
	Sig. (2-tailed)	0.86		0.24	0.37	0.90
	N	14	14	14	14	14
Passive	Pearson Correlation	0.12	0.33	1	-0.02	0.09
	Sig. (2-tailed)	0.67	0.24		0.93	0.74
	N	14	14	14	14	14
Visual	Pearson Correlation	**.61***	-0.25	-0.02	1	**.58***
	Sig. (2-tailed)	0.02	0.37	0.93		0.02
	N	14	14	14	14	14
Years Of Computer Use	Pearson Correlation	0.38	0.035	0.09	**.58***	1
	Sig. (2-tailed)	0.18	0.90	0.74	0.02	
	N	14	14	14	14	14
*. Correlation is significant at the 0.05 level (2-tailed).						

without it (t (10) = 2.294, p = 0.047). That was also enhanced by the results of the second experiment, which revealed that the mean student scores increased from (Mean = 9.94) to (Mean = 16.13) when they learn using the LAES system. Moreover, the results of a paired t-test revealed that there was a statistically significant difference between the mean scores of the participants when they learnt using the LAES and the mean participant scores when they learnt without it (t (16) = 2.289, p = 0.037).

More encouraging results emerged from the third experiment where the findings showed that the mean student scores increased from (Mean = 14.29) to (Mean = 22.14) when they learn using the LAES system. Moreover, the results of a paired t-test revealed that there was a statistically significant difference between the mean scores of the participants when they learnt using the system and the mean participant scores when they learnt without it (t (14) = -1.724, p = 0.048).

In general, the results revealed that the students had significantly higher learning outcomes when they used the LAES system to learn in a matched way. In addition, the (Cohen's d) effect size was medium (from 0.51 to 1.00) [20].

These results reject the first hypothesis (H_1), and prove that students who learnt using the ALSI system had significantly higher learning outcomes.

With reference to the second hypothesis, which is concerned with the correlation between dimensions of learning styles, the results were varied. While the first experiment revealed that there was a statistically positive significant correlation between visual and active styles, r (10) = 0.71, p = 0.02, the second experiment revealed that there was also a statistically positive significant correlation between passive and verbal styles, r (16) = 0.76, p = 0.00. Importantly, the results revealed that there was a statistically positive significant correlation between passive and active styles, r (16) = 0.79, p = 0.00. These results confirm that the dimensions of learning styles must not be treated as dichotomies (either/or options).

The results were also varied in terms of the third hypothesis, which is concerned with the correlation between the years of computer use and dimensions of learning styles. While the first two experiments revealed that there was no significant correlation between years of computer use and different learning styles, the third experiment revealed that there was a positive significant correlation between years of computer use and a visual learning style, r (14) = 0.58, p = 0.02.

References

[1] A. M. Alzain, S. Clark, G. Ireson and A. Jwaid, "LAES: An adaptive education system based on learners' learning styles." in The 2nd International Conference on Knowledge Engineering and Applications, 2017, pp. 107.

[2] A. Franzoni-Velázquez, F. Cervantes-Pérez and S. Assar, "A Quantitative Analysis of Student Learning Styles and Teacher Teachings Strategies in a Mexican Higher Education Institution," Journal of Applied Research and Technology, vol. 10, pp. 289-308, 2012.

[3] A. M. Alzain, S. Clark and G. Ireson, "Libyan higher education system, challenges and achievements," in Engineering Education (ICEED), 2014 IEEE 6th Conference On, 2014, pp. 67-72.

[4] Richard M. Felder And Joni Spurlin, "Reliability and Validity of the Index of Learning Styles: a Meta-analysis," International Journal of Engineering Education, vol. 21, pp. 103 - 112, 2005.

[5] A. L. Franzoni, S. Assar, B. Defude and J. Rojas, "Student learning styles adaptation method based on teaching strategies and electronic media," in Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference On, 2008, pp. 778-782.

[6] Z. A. Akasah and M. Alias, "Emphasizing learning of the affective domain for the realization of the engineering learning outcomes," Cognition, vol. 7, pp. 9, 2010.

[7] H. M. Truong, "Integrating learning styles and adaptive e-learning system: current developments, problems and opportunities," Comput. Hum. Behav., vol. 55, pp. 1185-1193, 2016.

[8] M. T. Alshammari, "Adaptation Based On Learning Style And Knowledge Level In E-Learning Systems, Ph.D thesis," 2016.

[9] P. Brusilovsky, "Adaptive hypermedia. User Modeling and User Adapted Interaction," Ten Year Anniversary Issue (Alfred Kobsa, Ed.), vol. 11, pp. 87 - 110, 2001.

[10] K. M. Feigh, M. C. Dorneich and C. C. Hayes, "Toward a characterization of adaptive systems: a framework for researchers and system designers," Hum. Factors, vol. 54, pp. 1008-1024, Dec, 2012.

[11] P. Brusilovsky, "Methods and techniques of adaptive hypermedia," in Adaptive Hypertext and HypermediaAnonymous Springer, 1998, pp. 1-43.

[12] N. Stash, "Incorporating cognitive/learning styles in a general-purpose adaptive hypermedia system," Dissertation Abstracts International, vol. 68, 2007.
[13] A. M. Alzain, G. Ireson, S. Clark and A. Jwaid, "Learning style instruments and impact of content: A qualitative study," in Sustainable Technologies (WCST), 2016 World Congress On, 2016, pp. 109-114.
[14] Ö Özyurt and H. Özyurt, "Learning style based individualized adaptive e-learning environments: Content analysis of the articles published from 2005 to 2014," Comput. Hum. Behav., vol. 52, pp. 349-358, 2015.
[15] S. Graf, T. Liu, N. Chen and S. J. Yang, "Learning styles and cognitive traits–Their relationship and its benefits in web-based educational systems," Comput. Hum. Behav., vol. 25, pp. 1280-1289, 2009.
[16] J. O. Liegle and T. N. Janicki, "The effect of learning styles on the navigation needs of Web-based learners," Comput. Hum. Behav., vol. 22, pp. 885-898, 2006.
[17] A. M. Alzain, S. Clark, G. Ireson and A. Jwaid, "A study of the reliability and validity of the first arabic learning styles instrument (ALSI)," in Sustainable Technologies (WCST), 2016 World Congress On, 2016, pp. 29-34.
[18] N. Al-Jojo, "Teacher assisting and subject adaptive material system: an Arabic adaptive learning environment, Ph.D. thesis," Teacher Assisting and Subject Adaptive Material System: An Arabic Adaptive Learning Environment, 2012.
[19] C. Mulwa, S. Lawless, M. Sharp and V. Wade, "The evaluation of adaptive and personalised information retrieval systems: a review," International Journal of Knowledge and Web Intelligence, vol. 2, pp. 138-156, 2011.
[20] L. Cohen, L. Manion and K. Morrison, Research Methods in Education. Routledge, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 109-115 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Designing Experiments: 3 Level Full Factorial Design and Variation of Processing Parameters Methods for Polymer Colors

Jamal Al Sadi*

Department of Mechanical Engineering, (HCT-ADMC). Shakhbout street, Abu Dhabi, UAE

A R T I C L E I N F O

Article history:
Received: 11 August, 2018
Accepted: 16 September, 2018
Online: 22 September, 2018

Keywords:
Color
Polycarbonate
Optimization
Processing
Trends
Characterization

A B S T R A C T

In this work, we investigate the effects of variation of processing parameters on the quality of dispersion of polycarbonate compound. In order to achieve appropriate pigments dispersion, we performed compounding process parameters optimizations, by investigating three processing parameters, temperature, screw speed, and feed rate. We utilized experimental design for the optimization of process parameters based upon three levels full factorial response surface methodology was utilized. The experimental designs, statistical and numerical optimization were performed using design expert software. Statistical equation was developed to understand individual parameters interactions on the values of color. The model was established as statistically significant based on diagnostic tests performed. Our analysis of variance (ANOVA) illustrates that the parameters of color (dL, da* and db*) are affected by the three investigated parameters. The process parameters required to attain color values in a minimum desirable deviation dE* of 0.8 were found through optimization to be equal to 245.26 °C, 741.27 rpm, and 24.72 kg/hr. Furthermore, we also demonstrate variations of the processing variables while other parameters remained constant (General Trends). Both strategies generated process parameters that were statistically significant.*

1. Introduction

Because color has a vital function for the production of polycarbonate pigments, materials need to be extruded with good dispersion properties and uniform particle sizes. In order to yield plastic with a commercialisable color, adding pigments to it is usually required. However, a great challenge is likely to be faced to attain the required color from the first attempt. Several variables affect the color properties of polymers compounding during their extrusion steps, including temperature, screw speed, feed rate, residence time, and screw configuration. Several researchers investigated the effects of such process variables on yielded color during polymers compounding [1- 2]. Being constituted of chemical species, it is likely that the pigments will take part in chemical reactions depending on process conditions. Thus, the correct selection of the right variables is vital to attain the color requirements. Furthermore, the time-temperature relationship can also affect the polymer characterestics. The required pigments dispersion and good uniformity can be attained by decreasing the viscosity of the resin and increasing the mixing time [3].

*Jamal Alsadi, (HCT-ADMC), Email: jalsadi@hct.ac.ae

Spectrophotometers can serve as important measures to control the quality, quantify color, and numerically compare variations in colors [3].

Allowable tolerance limits in particular terms of dL*, da*, db* or dE* are usually chosed by the client; however, for the polycarbonate grade-3 under this study, limits were equal or less ≤ 1.0 for dE or ≤ 0.6 for dL*, da*, db* [4]. The deviation in L*, a*, b* is represented as "dE*, where

$$dE^* = [(dL^{*2}) + (da^{*2}) + (db^{*2})] \quad (1)$$

Instead of using absolute values of color, color differences concerning target values regarding dL*, da* and db* are used. The total change in color, dE* is used to represent the color difference in the CIELAB color space [5].

Design of Experiments (DOE) is a planned approach that allows an experimenter to plan the experiments and determine cause-and-effect relationships. DOE is extensively used in numerous areas of science because it reduces the number of experiment that need to be performed.

For optimal dispersion of these pigments, the optimization of extrusion process parameters is required. Researchers designed experiments to evaluate the effect of process parameters on colour properties of a compounded polycarbonate grade. A regression model was generated. Several factors were contributed to color mismatch. Such factors need to be studied to understand their effect on output colour [6 ,7, 8].

Many experimental designs have been recognized as useful techniques to optimize process variables. A modified general factorial DOE has been employed for investigating the effect of changes in compounding process variables on gloss and surface appearance of a PVC sheet. [9] .Different types of RSM designs are available, including a factorial design, central composite design (CCD), Box-Behnken design, and D-optimal design. [10] The execution of a DOE involving the Box-Behnken design (BBD) has been reported to determine a relationship between processing parameters and viscosity variation for a wood-plastic compound [11].

The BBD, being a combined array design, requires fewer runs than Taguchi's crossed array designs and allows estimation for significant interactions. [12] It is the most efficient design in terms of runs and requires only three levels of each factor in order to generate a quadratic model. [12,13] To estimate curvature, other designs require either five levels of each factor such as in a central composite design (CCD) or even more experimental runs such as in a three level factorial design.

Analysis-of-variance (ANOVA) is essential to validate the significance and fitness of the model; it explains, whether the developed quadratic model is meaningful. It investigated the bearing of process parameters and interaction of these parameters. The robustness of RSM designs is ensured by considering the propagation of error (POE). POEs, a measure of the standard deviation of the transmitted variability in the output response, are caused by fluctuations in significant controllable process variables during experimentation assuming uncontrollable factors (noise) to be zero. [14].

In this study, an experimental investigation of the processing parameters was conducted using DOE. DOE was utilized to determine the optimum number of experiments to be run so that sufficient data was available for analysis. The designs were prepared for three processing parameters: temperature, speed, and feed rate. The effect of the processing parameters on output response parameters was studied. Experimentation for various grades was carried out to observe the effect of controlled variation of different processing parameters on the colour attributes of compounded plastics. The results were analyzed to determine an optimum set of processing parameters in order to ensure minimum wastages and timely delivery of orders.

Statistical Design of Experiments can be used to study the color responses to variation in these processing parameters with the help of methods such as the Response Surface Methodology (RSM). In this approach the first step is to properly design experiments in order to evaluate model parameters efficiently after performing experiments. Second step is to develop a second order polynomial for the responses [14].

$$y = \beta_o + \sum_{i=1}^{k} \beta_{ij} x_i + \sum_{i=1}^{k} \beta_{ij} x_i^2 + \sum \sum_{i<j=2}^{k} \beta_{ij}\, x_i x_j + \varepsilon \quad (2)$$

Where y is the predicted response, β_0 is a constant, β_i is the *i*th linear coefficient, β_{ii} is the *i*th quadratic coefficient, β_{ij} is the *i*th interaction coefficient, x_i is the independent variable, k is number of factors and ε is error. Coefficients of the model predicted through regression of the obtained experimental data. Details of parameter estimations for the model done by these authors are reported elsewhere [13]. RSM is a collection of statistical and mathematical techniques useful for developing, improving and optimizing process. The Three-level full factorial design is one of the most powerful and efficient experimental design among other response surface designs (central composite, Doehlert matrix, and Box Behnken designs). The ultimate aim of the present study is to employ the 3 level full factorial design to optimize the processing parameters to have a minimum deviation in color properties (dE* < 0.8)

Focus was extended to study the variations of independent of processing parameters. Parameters were usedare such as temp, speed, and feed rate were used to affect the dependent responses for consistent output color (L*,a*,b*,dE*). The procedure of controlling the variations of two Processing parameters and keeping the third parameter constant (general trends), focused on the variation of the optimal federates parameters to achieve a minimum desirable deviation dE* of 0.3.

Design optimization of the two procedures to precisely determine cause and effect relationships. Both designs yielded models that were statistically significant and optimal color were found.

2. Experimental Set up

Experimentation was carried out for the investigated material at Industrial Plant. A blend of two polycarbonate resins was used along with four different pigments, the color formulation of these grades in parts per hundred (PPH) is presented in Table 1. The melt flow index (MFI) for Resin 1 was 25 gm/10min, and that of Resin 2 was 6.5 gm/10min.

Table 1: Color formulation used for investigated grade

S.No	Type	PPH
1	Resin 1	33
2	Resin 2	67
3	Pigment A	0.20
4	Pigment B	0.05
5	Pigment C	0.0004
6	Pigment D	0.0016
7	Pigemnt E	0.0710

3. Design of Experiments

The Design of Experiments (DOE) containing the 27 experimental runs was used to implement a Three-Level Full Factorial Design and a DOE containing 9 experimental runs was used to implement variations of the processing parameters while other parameters are constant (General Trends) are shown in Table 2 and Table 3 respectively .

Table 2: Design level in actual and coded unit

Parameters	Units	3 Levels		
		-1	0	+1
Temperature	ºC	230	255	280
Speed	rpm	700	750	800
Feed rate	kg/h	20	25	30

3.1 3 Level Full Factorial Design

The processing was carried out on a twin- screw extruder of 25.5 mm diameter, with ratios of L/D=37 and D_o/D_i=1.55. The materials were extruded in an intermeshing , ZSK26- Coperion - Germany ,27 kW twin co rotating screw extruder (TSE). The three process parameters, the temperature of the heating zones, feed rate to the extruder, and the screw speed were considered in the experimental design, and the levels used are shown in Table 2. Parameters were varied on 27 different treatments with additional five center points, the total of runs are (32 treatments) for 3 level full factorial response method to study their effects on color. The additional five centre points were added to estimate the experimental error and for the detection of nonlinearity in the responses [13].

Extruded melt was quenched in cold water, air dried and then pelletized. Using injection molding, the pellets from each run were molded into three rectangular chips (3x2x0.1") after which their values (CIE L*, a*, b*) were investigated by utilizing a spectrophotometer (CE 7000A, X rite- Inc. USA). Target color output for these values were L*= 70.04, a*= 3.41, and b*= 18.09. Statistical analysis of data was performed using The Design Expert Software (Version 8, Stat-Ease Inc. USA) to quantify and relate the effects of variables at a confidence interval of 95%. In order to attain zero deviation from target color, numerical optimization of the data was performed.

3.2 Variations of the Processing Parameters (General Trends)

Because color is directly related to the process parameters involved, herein, we performed a control study to investigate the effects of operating variables (temperature, speed, and feed rate) on color. Three processing parameters were controlled individually at three different stages, while fixing all other parameters (GT). Based on our observed strong correlations between the processing variables and the color generated, we conclude the following recommenations: Flow rate was 20 kg/hr, 25 kg/hr, and 30 kg/hr, at a speed of 750 rpm and a temperature of 255 °C [15-17] .

The selected processing temperatures were 230°C, 255°C and 280°C with a speed and flow rate fixed at the middle values (750 rpm and 25 kg/hr, respectively). A similar procedure was followed for both speed and flow rate. The selected speeds were 700, 750 and 800 rpm and the selected flow rates were 20, 25, and 30 kg/hr. The following tables show the experimental processing conditions. The general trends (GT) experimental design is shown in Table 3.

Assuming that the aforementioned variables were utilized, in this work we suggest optimized process parmeters to attain plastic grade color consistency.

Table 3: Processing Parameters Variables

Speed RPM	BZ1 (°C)	BZ2 (°C)	BZ3 (°C)	BZ4 (°C)	BZ5 (°C)	BZ6 (°C)	BZ7 (°C)	BZ8 (°C)	BZ9 (°C)	DZ1 (°C)	F. R Kg /hr
750	70	195	230	230	230	230	230	230	230	230	25
750	70	195	255	255	255	255	255	255	255	255	25
750	70	195	280	280	280	280	280	280	280	280	25
700	70	195	255	255	255	255	255	255	255	255	25
750	70	195	255	255	255	255	255	255	255	255	25
800	70	195	255	255	255	255	255	255	255	255	25
750	70	195	255	255	255	255	255	255	255	255	20
750	70	195	255	255	255	255	255	255	255	255	25
750	70	195	255	255	255	255	255	255	255	255	30

4. Results

Using analysis of variance (ANOVA), correlations between variables were investigated and processing parameters were optimized to generate resulting color properties. The Design Expert software was used to the effects of the operating process parmeters on dL*, da*, db*. We performed sequential F-tests, by utilizing an initial linear model, modified by subsequently adding suitable linear or quadratic terms [20]. The F-statistic was considered for each type of model, and the highest order model with significant terms was chosen. Analysis of all tristimulus values was based on the same process. Based on the F statistics ≤ 0.05 and probability values ≤ 0.1, only significant terms were added. Our ANOVA sequential model sum of squares results for dL*, da*, db* are given in Table 4.

The highest order model with significant terms (*Prob* > *F* is less than 0.05) are the 2F models, and are applicable to describe the dL*, da* and db* responses. R^2 values (Table 4) provide confirmation and suggest that a variability of 78% in dL*, da* and db* is decreased to 74 %. The remainng variability is unexplained and can be ascribed to noise. The "Predicted R-Squared" and the "Adjacent R-Squared" are in rational agreement. A signal to noise ratio larger than 4 is usually desirable since it suggests that the model can be utilized for navigation into the design space, and can be quantified by "Adequate Precision". The "Lack of the fit" test was utilized to compare between the residual and the pure error, and resembled a p-value > 0.05, which is insignificant.

Table 4: ANOVA results for dL*, da* and db* response

Response	Significant Terms	R^2	Predicted R^2	Adjacent R^2	Adequate Precision
dL*	A, B, C , AB, AC, BC	0.78	0.38	0.55	9.70
da*	B, C, BC	0.75	0.24	0.39	8.53
db*	B,C	0.75	0.28	0.30	8.61

To generate predicted responses for the dL*, da* db*, several models based on linear regression were produced as given in Table 5. The effects of temperature, speed, and feed rate operating variables and their interactions are represented by the polynomial equations. The effect of these variables on the responses correlate with the coefficient values. The differnce between predicted and actual values is in the 0.2 limit suggesting a strong agreement between these 2 values.

RSM curves were used to investigate interactions and optimize process parameters variables. The contour graph (Figure 1a) illustrates the first order relation between temperature and speed, at the feed rate of 24.71kg/hr for dL*, and resembles that various temperature and speed combinations can satisfy the required objective. Moreover, the contour graph (Figure 1b) illustrates the relation between temperature and the feed rate of 741.2 rpm. Finally, the contour graph (Figure 1c) illustrates the relation between speed and feed rate at temperature of 245.2 °C. The global optimal value (dL* = 0.0) is realized at 245.2 °C, 741.2 rpm and 24.71 kg/hr as demonstrated in the graphs. The maximum and minmum dL* values at 95% confidence interval are both equal to 0.11.

Table 5: Regression Model for dL*, da* and db*

Response	Regression model
dL*	+63.86390 - 0.19647 * Temperature -0.065085 * Speed - 0.99472 * Feed Rate +1.84353E^{-004} * Temperature * Speed +1.96624E^{-003} * Temperature * Feed Rate + 6.39611E^{-004} * Speed * Feed Rate
da*	+14.59778 - 0.018496 * Speed - 0.47296 * Feed Rate + 5.98224E^{-004} * Speed * Feed Rate
db*	+4.08697 - 4.78866E $^{-003}$ * Speed - 0.029746 * Feed Rate

A= Temperature, B= Speed, C = Feed rate

Table 6: dL*, db* and da* Actual and Predicted Values

Color	dL*		da*		db*	
Run NOs	Actual Value	Pred. Value	Actual Value	Pred. Value	Actual Value	Pred. value
1	-0.52	-0.5	0.61	0.57	0.31	0.14
2	0.16	0.093	0.23	0.12	-0.17	-0.25
3	-0.59	-0.57	-0.3	-0.087	-0.76	-0.34
4	0.11	-0.2	0.29	0.003	-0.22	-0.4
5	0.01	-0.24	-0.16	-0.059	-0.16	-0.49
7	-0.65	-0.36	-0.29	0.12	-0.77	-0.25
8	0.14	-0.029	0.34	0.3	-0.11	0.008
9	-0.53	-0.47	0.027	0.024	-0.38	-0.16
10	0.087	-0.18	0.14	0.03	-0.18	-0.4
11	-0.21	-0.087	0.077	-0.087	-0.56	-0.34
12	-0.44	-0.54	0.65	0.24	0.37	-0.099
13	0.25	0.41	0.27	0.24	-0.14	-0.099
14	-0.55	-0.23	-0.15	-0.059	-0.55	-0.49
15	-0.71	-0.2	-0.05	0.031	-0.93	-0.4
16	-0.53	-0.4	-0.12	-0.031	-0.84	-0.64
18	-0.34	-0.26	-0.22	0.024	-0.35	-0.16
19	0.037	-0.065	0.29	0.24	-0.22	-0.099
20	-0.48	-0.49	0.65	0.3	0.37	0.008
21	-0.2	-0.33	-0.017	-0.087	-0.33	-0.34
22	0.43	0.43	0.32	0.3	0.013	0.008
23	0.027	-0.13	0.24	0.12	-0.24	-0.25
24	0.087	0.2	0.24	0.57	-0.16	0.14
27	-0.04	-0.13	0.05	0.12	-0.31	-0.25
28	-0.15	-0.13	0.12	0.12	-0.013	-0.25
29	-0.19	-0.13	0.093	0.12	-0.083	-0.25
30	-0.02	-0.13	-0.087	0.12	-0.02	-0.25
31	-0.1	-0.13	-0.09	0.12	-0.06	-0.25
32	-0.033	-0.15	0.13	-0.031	-0.32	-0.64

Figure 2 shows the interaction between speed and feed rate for da* at 245.2 °C, resembling the first order relationship and suggest that several feed rate and speed combinations can satisfy the objective. The optimum value of da* = 0.15 is realized at 741.2 rpm and 24.7 kg/hr. The maximum and minimum allowable values at 95% confidence interval for da* are 0.23 and 0.07 respectively.

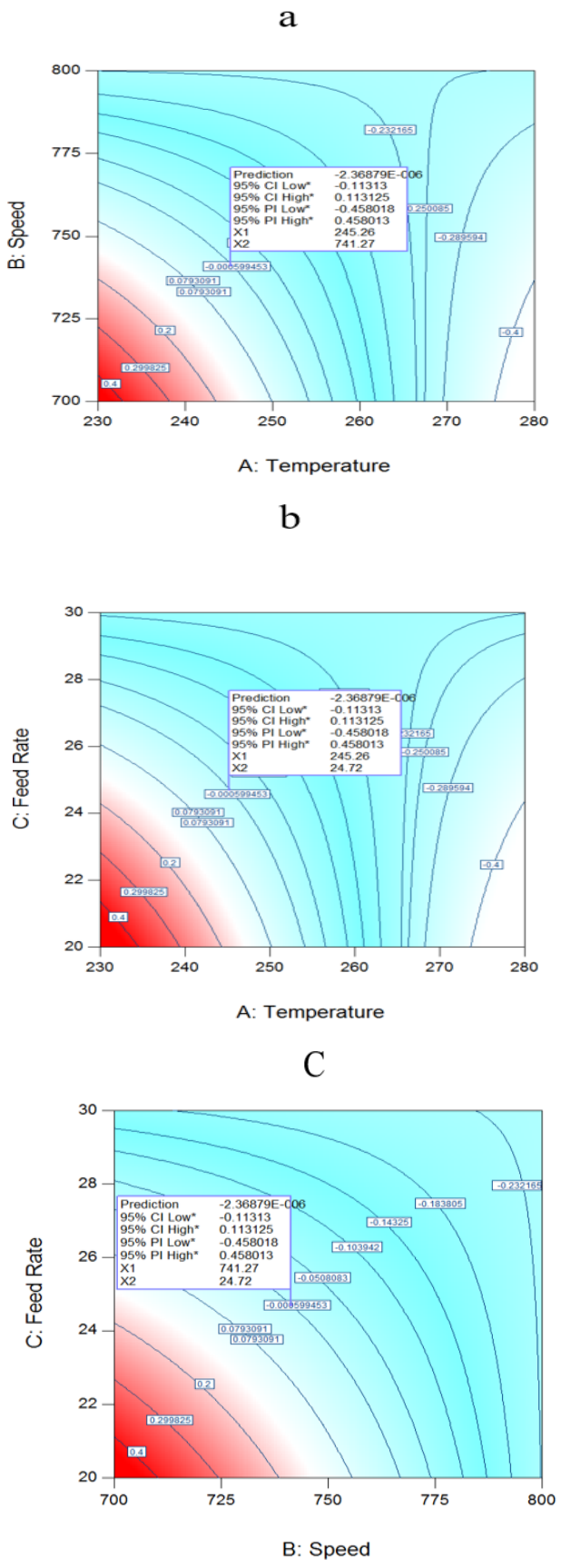


Figure 1: (a) Interaction between temp and speed at 24.7 kg/hr for dL* (b) Temp and feed rate at 741.2 rpm for dL* (c) Speed and feed rate at 245.2 °C for dL*.

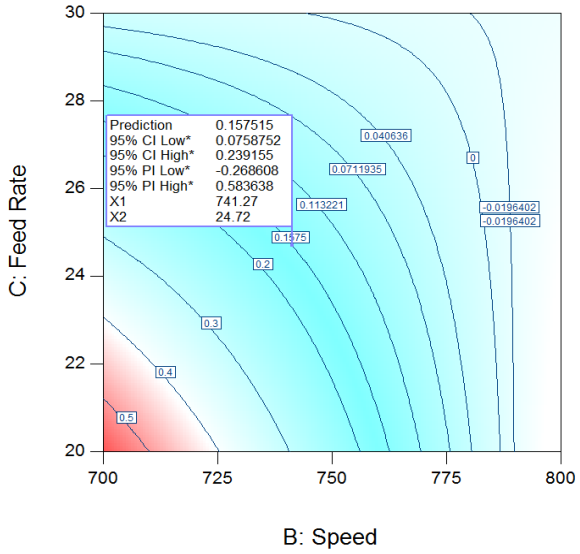


Figure. 2: Speed and feed rate interaction at temperature of 245.2. ºC for da *

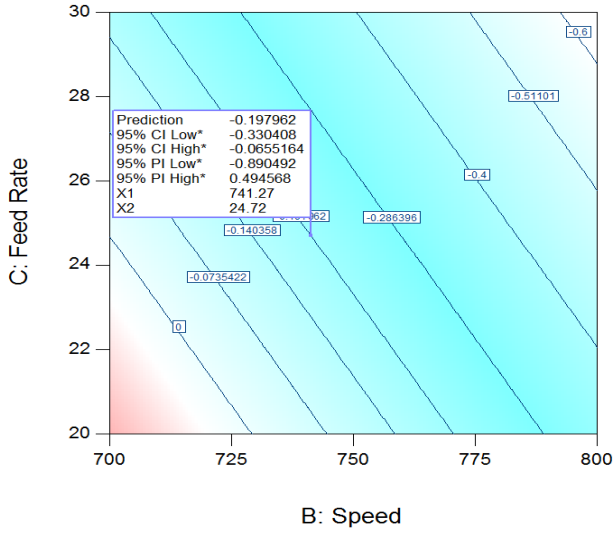


Figure 3: Speed and feed rate contour plot for d b* at 245.2 ºC

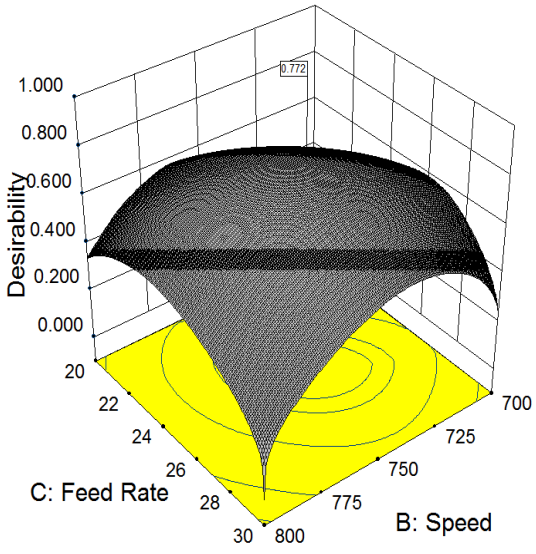


Figure 4: Desirability Graph at temperature of 245.2 ºC (A= Temp; B= Speed, C= Feed rate)

The interaction between speed and feed rate for db* at 245.2 ºC is shown in Figure 3. The linear behavior includes only speed and feed rate as significant model terms as compared to dL* and da*. Interestingly, at lower values of speed and feed rate around 20 kg/hr and as well as at higher values of speed and feed rate around 24 kg/hr, db* approaches zero, while temperature stays at 245.2 ºC. At the global optimum, the predicted value of db* is equal to -0.19 and the allowable maximum and minimum values at 95% confidence interval are -0.05 and -0.33 respectively.

An important effect on the color responses was exerted by temperature (A), speed (B) and feed rate (C). The optimal settings of the three parameters regarding all responses were realised by utilizing a decision making method based on multiple criteria, and a total desirability function "d" [19]. A desirability function is quantifies the quality and presents a convenient responses comparison method to select the optimal settings; Figure 4 shows a 3D plot of the global desirability *D*, maintaining a feed rate of 24.7 kg/hr, and indicates that the maximum combined desirability of 77% is attained at 245.2ºC, 741.2 rpm, and 24.7 kg/hr.

5. Effect of variation feed rate on color values

A plot of feed rate variation from 20 kg/hr to 30 kg/hr with a fixed speed at 750 kg/hr and a fixed temperature at 255 °C is shown in Figure 5; The difference in color decreases initially as the feed rate increases up to 30 kg/hr.[20-22]

A response contours plot of speed versus feed rate at a temperature of 245.2 ºC is shown in Figure 6, and resembles the presence of region that is feasible to achieve target values. The plot shows a region between the da*=0.30 and db*=0.20 contours which illustrates the temperature and speed operation conditions at which the mean responses (dL*, da*, db* and SME) target are met at a fixed feed rate of 24.44 kg/hr.

The results suggest that the optimal tristimulus values of dL* = 0.0, da * = 0.15, db* = - 0.19 are attained at 245.2 ºC, 741.2 rpm, and 22.7 kg/hr. The total minimum deviation in tristimulus values (using equation 1) is equal to 0.25, which is reasonably acceptable as compared with the maximum allowable deviation (dE* = 0.8).

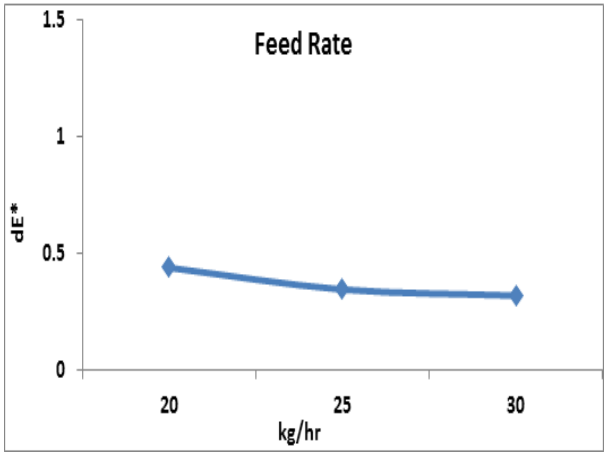


Figure 5: Effect of the variation of feed rate at fixed a fixed temperature of 255 °C, and a fixed speed of 750 rpm.

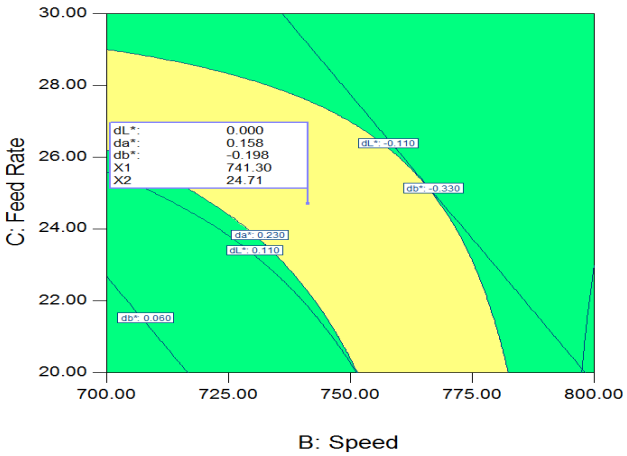


Figure. 6: Speed and feed rate overlay plot at a temperature of 245.2 ºC.

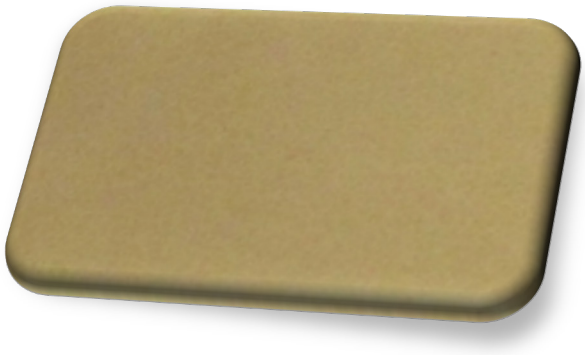


Figure 7 : Rectangular color chip (3x2x0.1")

Using the lab injection molding, the extruded pellets were molded into a rectangular chip (3x2x0.1") after which the color values (CIE L*, a*, b*) were examined by using a spectrophotometer (CE 7000A, X rite- Inc. USA). The image of the specimen is shown in Figure 7.

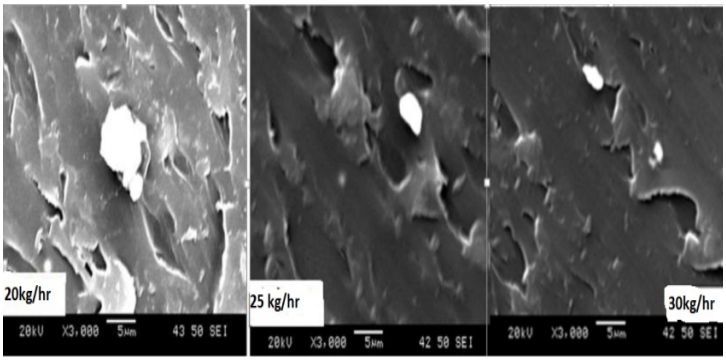


Figure 8: Scanning Electron Microscopy images at feed rates of (a) 20 kg/hr, (b) 25 kg/hr, and (c) 30 kg/hr, at a fixed temperature of 255 °C, and a fixed speed of 750 rpm.

Scanning Electron Microscopy (SEM) images were probed using a Joel 5500 LV at a 20 kV voltage, at different feed rates, and are shown in Figure 8. The images illustrate that primary particles and agglomerates were present in the compounded grades. The images show that larger agglomerated pigments could be observed at a lower feed rate as compared with higher feed rates [20-22].

Conclusions

An experimental design approach which relied on three levels full-factorial surface method was utilized to study the effects of operational process variables, and suggested appropriate predictive models for dL*, da*, and db*. We used a regression model to calculate tristimulus values and validated our work by confiming that a good agreement exists with experimental results. At target values of 245.2 ºC, 741.2 rpm, and 24.7 kg/hr, tristimulus values were close, with a minimum total deviation (dE*) of 0.25 at a 95% confidence interval. An experimental design approach, which relied on the interaction between tristimulus color values and processing parameters showed that the minimum color difference throughout the experiment was equal to 0.34 for dE*. As the feed rates increased from 20 kg/hr to 30 kg/hr, the color difference values (dE*) substantially decreased. Finally, an optimum set of processing parameters for the grades of the polycarbonate can be yielded through the utilization of the optimized parameters, and hence reduce colors mismatch so that wastages can be reduced.

References

[1] S. P. Rwei, Distributive Mixing in a Single-Screw Extruder – Evaluation in the Flow Direction, Polymer Engineering and Science, (2001).
[2] A.Y. Wong, Screw Configuration Effects on The Color Mixing Characteristics Of Polymer In Single-Screw Extrusion, The University of Hong Kong, China Tinhua Liu, Sichuan Union University, China (1998).
[3] D. I. Meade, "Introduction to colorant selection and Application Technology," in Coloring of Plastics, Fundamentals, 2nd ed., edited by Robert Charvat (2004).
[4] ASTM D 2244-93. Standard Test Method for Calculation of Color Differences from Instirnentally Measured Color Coordinates. 222-215
[5] Richard Abrams, Plastic additives, and compounding (2001)
[6] Rauwendaal, C., Polymer Mixing, A Self-Study Guide. Hanser Publishers, Munich., 1998.
[7] Rwei, S.P., Distributive Mixing in a Single-Screw Extruder – Evaluation in the Flow Direction. Polymer Engineering and Science, 2001
[8] Mulholland, B.M., Effect of Additives on the Color & Appearance of Plastics. SPE ANTEC. Ticona, Ticona Engineering Polymers ,2007.
[9] Effertz, K., Understanding the Effects of a Compounding Process on the Production of Co-Extruded Vinyl Sheet through the Utilization of Design of Experiments (part II). ANTEC-Conference Proceedings: p. 3574-3578 ,2004
[10] http://itl.nist.gov/div898/handbook. [Online] available accessed in 2013.
[11] Bender, T.M., Characterization of Apparent Viscosity with respect to a PVC-Wood Fiber Extrusion Process. ANTEC-Conference Proceedings, 2002.
[12] Borror, C.N., Montgomery, D. C. and Myers, R. H., Evaluation of Statistical Designs for Experiments Involving Noise Variables. Journal of Quality Technology. Vol 34(1): p. 54-70 , 2002.
[13] Anderson, M.J., and Whitcomb, P. J., RSM Simplified: Optimizing Processes using Response Surface Methods for Design of Experiments. New York, NY: CRC Press. , 2005.
[14] Montgomery, Douglas C., Design and Analysis of Experiments, 6th Ed., John Wiley and Sons, New York,2005.
[15] J. AlSadi1, S. Ahmad1, U. Saeed1, G. Rizvi1, D. Ross2, R. Clarke2, J. Price2. (2012) Execution of 3 level full factorial design to evaluate the process parameters: polymer color properties, Annual Technical conference of the Society of Plastics Engineers (ANTEC), Orlando, Florida, USA, pp. 1-5, 2012.
[16] Saeed.U. AlSadi, .J, Ahmed. S, Rizvi. G., Ross. D, 2013, "Neural Network: a potential approach for error reduction in color values of polycarbonate" Journal Advance polymer technology. 33 (2), Summer 2013, DOI 10.1002/Adv. 21402
[17] J. AlSadi, M. Rabbani, S. Ahmed, G. Rizvi, R. Clarke and D. Ross, (2011) Effect of Processing Parameters on Colour During Compounding, Annual Technical Conference of the Society of Plastics Engineers (ANTEC), Boston, USA, pp. 1-4, 2011
[18] Y. K. Chang, M. K. Bustos, H.Lara, Braz. J. Chem. Eng. vol. 15 n. 4 São Paulo Dec. 1998
[19] M. A. Islam, V. Sakkas, T. A. Albani, J. Hazardous Materials, (170) 2009.

[20] J. Alsadi, U. Saeed, S. Ahmad, G. Rizvi and D. Ross , Processing issues of color mismatch: Rheological characterization of polycarbonate blends, Polymer Engineering & Science, Volume 55, Issue 9, September 2015, Pages: 1994–2001, Version of Record online : 8 DEC 2014, DOI: 10.1002/pen.24041

[21] AlSadi, J., Ahmad, S. , Saeed, U. , Rizvi, G. , Ross, D. , Clarke, R. and Price, J., Effects of Processing Parameters on Color Mismatch During Compounding. Annual Technical conference of the Society of Plastics Engineers (ANTEC), Orlando, Florida, USA ,pp 1-5 , 2012.

[22] J, Alsadi . Colour Mismatch In Compounding Of Plastics: Processing Issues and Rheological Effects (Doctor of Philosophy in Mechanical Engineering), The Faculty of Engineering and Applied Science University of Ontario Institute of Technology-UOIT). Ontario, Canada(2015).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 116-124 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

An enhanced Biometric-based Face Recognition System using Genetic and CRO Algorithms

Ola Surakhi*,1, Mohammad Khanafseh[1] , Yasser Jaffal[2]

[1]*University of Jordan, King Abdullah II School for Information Technology, Computer Science Department,11942, Jordan*

[2]*University of Kassel, Department of Technical Computer Science, Germany*

ARTICLE INFO

Article history:
Received: 04 May, 2018
Accepted: 31 May, 2018
Online: 15 June, 2018

Keywords:
Face Recognition
Biometrics-based security
Genetic algorithm
CRO algorithm
Optimization

ABSTRACT

Face recognition is one of the most well-known biometric methods. It is a technique used for identifying individual from his face. The recognition process takes the face and compares it with the one stored in the database for recognizing it. Many methods were proposed to achieve that. In this paper, a new technique is proposed by using meta-heuristic algorithm. The algorithm is used to search for the best point in the image to be selected as a pivot, generate a vector of extracted features that are not necessary for the recognition and may reduce accuracy of it and evaluate the weight value for each area in the face image areas. A dataset with 371 images was used for evaluation. The results were conducted and compared with the original face recognition technique. That results show that proposed idea could enhance recognition by increasing accuracy up to 20% over original face recognition technique.

1. Introduction

Recently, Face recognition has received a great attention in many fields from security, psychology, and image processing, to computer vision [1-5] because of its accuracy and low intrusiveness. It is one of the biometric techniques used for identification individuals by comparing face image with the ones stored in the database [6]. The process of face recognition can be divided into three stages: face detection, feature extraction and face recognition as shown in Figure 1



Face detection detects the existence of the face in an image by locating the position of it. Then, the features are extracted which is the most important step to recognize face. The face features are extracted to a vector, which is considered as a signature to the image. The features for each face are unique and used to discriminate between two individuals. Last, face recognition involves verification and identification. Verification compares the face image to approve the request of individual, identification compares face image with the set of images stored in the database to identify it.

Many methods had been proposed before for face recognition such that Principal Component Analysis (PCA), Multi-linear Principal Component Analysis (MPCA) and Linear Discriminate Analysis (LDA) [8,9]. In [10], the authors proposed a new face recognition using genetic algorithm. The work compared with the Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) algorithms for face recognition and analyzed the face recognition results using various databases such as ORL, UMIST and Indbase. The results clearly show that the recognition rate of Genetic algorithm is better than the PCA and LDA in case of ORL, UMIST and Indbase databases. In [11], the authors proposed a novel algorithm for recognizing human faces using Genetic Algorithms. The proposed method produces better result as compared to other techniques like PCA and LDA for one sample per person.

1.1. Face recognition system

For each image, there are more than 80 feature points on human face and one of these feature points must be elected as

* Ola Surakhi, 1University of Jordan, Email: ola.surakhi@gmail.com

pivot point. The pivot point is the most important feature point, where the distance between it and all other feature points must be calculated and saved into database to be compared lately with new results of user who needs to enter to the system. Based on this calculation and features, the idea of original face recognition goes through two steps, enrollment and matching phase as shown in Figures 2 and 3.

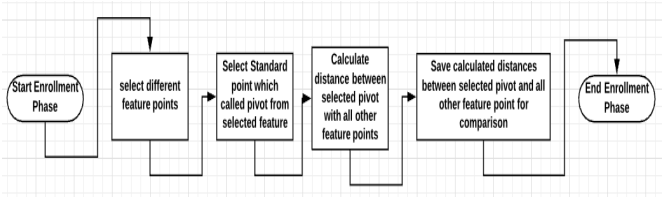


Figure 2: Face recognition enrollment phase

After enrollment phase, all the user's data are stored now in the database. Next step is the matching phase. Through this phase, the distance between pivot point and all other feature points will be calculated for the user who needs to access. The output result from this calculation will be compared with the data stored in the database. If the number of matched feature distance is greater than a threshold value then the result for this user is matched, else this user is not matched.

The human face can be divided into different areas such as eyes area, nose area, mouth area, etc. Each of these areas has a specific weight value; this weight value plays an important role in matching result. The flowchart for the matching phase is shown in Figure 3.

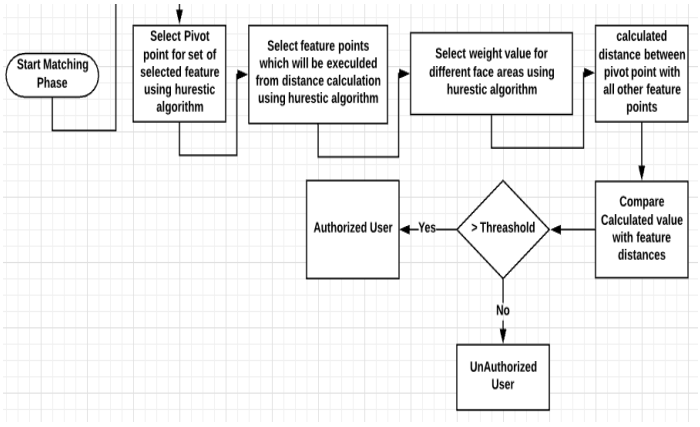


Figure 3: Face recognition matching phase

Face recognition technique show that there are different factors that can play an important role in matching accuracy. Based on these factors, we propose a new scheme for face recognition to achieve high level of accuracy that depends on using a meta heuristic algorithm to select best values for different variables; two meta heuristic algorithm used in our approach: Genetic meta heuristic algorithm (GA) and chemical reaction optimization meta heuristic algorithm (CRO). The goals of using them can be summarized as follows:

1. Select the best point in the image to be a pivot. The pivot represents a point in the image where the distance between it and every point in the features can be used for enhancing recognition rate.
2. Select a set of features to be excluded from the processing. Some of the extracted features in the image can make noise and decrease accuracy of detection; for that the proposed algorithm used a vector to store such features to be excluded from the recognition process.
3. Update weight values for each face area to enhance its accuracy.

1.2. Applications of Face recognition system

Nowadays, many applications in many areas that need personal identification use face recognition techniques; some of these are listed below:

a) Access control system: It uses face as a verification method to match with the one enrolled.
b) Banking
c) Internet and e-commerce
d) Human computer interface HCI
e) Smart cards
f) Security camera system presents common in airports, companies, universities, ATM machines, among others.
g) Replacement of PIN
h) Others

1.3. Advantages of using Face recognition system

Using biometric techniques in general has a lot of advantages. Face recognition method is most widely used because of the following advantages:

1. The existence of camera in most wireless devices
2. Uniqueness
3. Inexpensive technique
4. The system does not require any direct communication between user and device in order to authenticate him/her.

1.4. Disadvantages of Face recognition system

There are a number of disadvantages of using face recognition techniques, which can be summarized as follows:

1. Aging: The picture of the individual may be change by age and need to be updated in the database every two years or so in average.
2. Using of special characters such as glasses may affect on the accuracy of the recognition.
3. Emotion of the individual such as sadness, happiness, and anger will have an influence on the picture.
4. Image pose: The camera pose will affect the image as well as the rotation of the face.

2. Proposed work

2.1. Dataset

The dataset used in this paper is taken from XM2VTSDB multi-modal face database project [12], which consists of 371 persons picture features with different sessions for each person. The overall number of images collected on the database is about

2360 different images for 371 persons. From each image we extracted 67 features. These collected images refer to different genders, males and females, with different ages and for different positions for each person. These images are taken over a period of four months.

2.2. Genetic algorithms

A Genetic algorithm, GA [13,14,15], is defined as the most powerful algorithm used to solve large problems where there are a large number of solutions and an optimization is needed to generate better solutions from them. The main components of the algorithm are:

1. An initial population for the solution
2. A fitness function to generate better solution

Each individual in the population is a numeric value with n bits that represents a solution to the problem. The individual in the GA is called chromosome, which is a sequence of 1's and 0's. In order to generate better solution from the population, GA runs a set of operations to optimize solution and gets better results for the problem as follows:

1. Evaluate fitness function for each individual.
2. Selection: The selection operator selects individuals from the population with highest fitness value as a parent.
3. Crossover: This is the recombination operator, which allows individuals to exchange information and generate the offspring for the next population. The aim is to generate better offspring who is not identical to his parents but contains their traits. Three types of crossover can be used: one point, two point and uniform.
4. Mutation: Is making a change in some bits value of the chromosome in order to generate new better individual.
5. Update value of the fitness function for each individual.

All the above three steps are repeated to a maximum number of iterations.

2.3. Chemical Reaction Optimization (CRO) Algorithm

Chemical Reaction Optimization (CRO) is a meta-heuristic algorithm inspired by the nature of chemical reactions [16]. It starts with initial molecules, and by doing a sequence of collisions where final product becomes in the stable state. The major difference between CRO algorithm and any other evolutionary techniques is that population size in the CRO may change after each iteration of the algorithm running, while in all other techniques, the population size is fixed and unchanged during execution. The basic unit in the CRO population is called molecule, which has a potential energy that is considered fitness function of the individual. Each molecule has a set of parameters like kinetic energy, molecule structure and more, some of them are important and some are less important depending on the problem.

In order to change on the molecule, a collision is made which could be either uni-molecule (one molecule) or inter-molecule (two or more collide with each other). The aim is to transform into a stable product with minimal potential energy. The chemical reaction could be one of these types:

a. On-wall ineffective collision: This occurs when the molecule collides with the wall of container and then bounces. The transformation of the molecule structure can be represented as ω → ω-
b. Decomposition: This happens when the molecule hits in the wall and then decomposes into small parts. ω → ω1- + ω2
c. Inter-molecular ineffective collision: y happens when multi molecule collide with each other and then bounces away.
d. Synthesis: It is the opposite to the decomposition and happens when two or more molecules hit and combine together; ω1 + ω2 → ω-.

The steps of the algorithm can be summarized as shown below and the pseudo code is shown in Figure 4.

1. Starts with the initial population, which consists of a set of individuals where each one has a potential energy (PE). Some of the CRO parameters such that population size, number of iterations and buffer should be defined initially.
2. Apply chemical reaction to generate new reactants.
3. Update the potential energy.
4. Repeat steps till reaching termination condition.

```
1) begin
2) initialization
3) judge rand()>MoleColl
4) 3)satisfied, then judge KE ≤ β
5) 4)satisfied, synthesis
6) inter-molecular ineffective collisions
7) 4) didn't satisfied, synthesis
8) 3) didn't satisfied, molecule selection
9) judge NumHit − MinHit > α
10) 10) satisfied, on-wall ineffective
11) decomposition
12) check for min PE
13) curFE<parameFElimit satisfied, return 3)
14) end
```

Figure 4: pseudo code of the CRO algorithm

2.4. Proposed scheme

This paper proposes an enhancement version over available face recognition schemes using a meta-heuristic algorithm to achieve a high level of accuracy and less error-matching rate. The original face recognition system consists of two phases: enrollment and matching. The system depends mainly on choosing a point as a pivot point and then calculate the distance between it and other face features. The data is then stored as a vector in the database for each image. The second phase compares the values of the registered user with the data stored in the database for matching and authenticating individual.

Our proposed scheme is implemented using two meta heuristic algorithms: first one is the improved face recognition using Genetic algorithm and second deals with a new version of face recognition using chemical reaction optimization algorithm. The goal of both algorithms is to enhance both accuracy level and decrease error rate of the matching process. The results from each one is conducted and compared to show which algorithm enhances accuracy and gives better results.

2.4.1 An improved face recognition technique using Genetic algorithm

Face recognition techniques depend on two main things: (a) finding pivot point to calculate distance between selected pivot and all other feature points and (b) dividing human face into different areas and assigning a specific weight value for each area, the area which have highest weight value will be the most important area which influences on matching process by reducing error rate of matching, especially when the individual image has some special characters like using glasses or mustache, in such cases the area with glasses will be assigned a low weight value so it will not affect that much on the accuracy of the recognition. An example of the image features used in the training is shown in Figure 5.

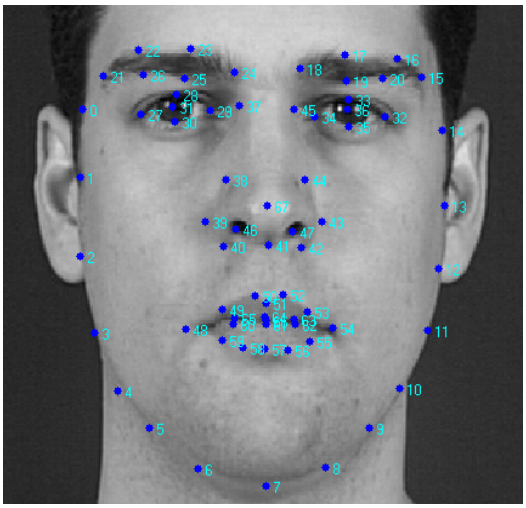


Figure 5: Example of image features

The GA is used first to select pivot point. Second, some of the feature can be excluded into an exclude array based on the nature of the image; the GA is used to change this array contents till reaching to the highest matching result and decreasing error rate. Third, GA is used to select best weight value for each of the face areas such that total weights for all must be equal to100%.

Genetic algorithm consists of different main phases as we explain previously: the fitness function here will be the matching value between entered image input and the one stored in the database.

Each of genetic main phases has a special work on our proposed idea. The following table presents different genetic algorithm phases and their mapping in our scheme.

Table 1: Mapping between different genetic phases and our proposed scheme to find heights match value

Genetic Phase	Its meaning on our proposed idea
Individual	Pivot point, weight for each face area, set of excluded features points and distances from selected pivot and all feature points
Population	Set of individuals
Search Space	Different solutions founded through different iterations.
Fitness Function	Match values for testing data sets based on training data for all faces, best solution will have highest fitness value which mean highest match rate for different images which compared with all feature information saved on DB.
Crossover	Generate different values for pivot, face area weight, excluded array based on best solution with other solutions
Mutation	Random difference for generated solution based on specific value

First solution will be generated randomly; different parameters will be generated from this solution such as selecting a pivot value, selecting weight for each of face area and set of excluded features to be avoided. All these values will be defined and generated through initialization step of genetic algorithm in order to build other solutions based on them. The fitness value is then evaluated and assigned for each individual in the set of solutions. Based on the fitness value the selection step starts.

Second step of genetic algorithm refers to selection step. Select a set of individuals from population with highest fitness value to be parents which will produce a new offspring with better fitness value.

Third step of genetic algorithm is crossover step; apply crossover between selected solutions and other solutions in the population in order to define a set of good solutions with highest match value (highest fitness value) as possible.

Last step of genetic algorithm refers to mutation step. This step is implemented to make changes in the new generation by multiplying it with a ratio value. This ration must be small between 0.01 and 0.025 of all population to change the solution.

After finishing all of these steps, a new solution will be generated with highest matching value and more accurate than original solution.

2.4.2 Face Recognition Using Chemical Reaction Optimization (CRO) Algorithm

In this section, the Chemical Reaction Optimization (CRO) meta heuristic algorithm is used for face recognition technique. CRO will be used to choose variables: pivot point, excluded feature points and different weights for face areas such as what we did when used GA. The main idea is to use CRO algorithm to select values for these important variables which influence on

matching results. CRO generates better solution after each iteration till reaching best one. Table 2 presents some of CRO scheme meanings and their mapping to our proposed technique.

Table 2: Mapping between chemical meaning and its meaning on face recognition.

Chemical Meaning	**Its meaning on our proposed idea**
Molecular Structure	Set of Solutions which found based on original solution
Potential Energy	Important variables value: Pivot value, Exclude array values, Weights value for different face areas
Kinetic Energy	Measure of tolerance to have a worse Solution
Number of Hits	Total Number of iteration used for specific experiment.
Minimum Structure	Current Optimal value for Matching based on Different variables values
Synthesis Interaction, $\omega1 + \omega2 \rightarrow \omega'$	Two solutions with two Potential Energy combined with each other to select single solution with highest Potential energy which refers to highest match percent for all images.
Inter-molecule infective collision $\omega1 + \omega2 \rightarrow \omega1' + \omega2'$.	Two solutions with two Potential energy will produce a single solution with highest Potential energy value. By combining different steps of both solutions as select best exclude array values from one solution with face area weight from other solution to have a solution with highest match percent for all faces.
Decomposition $\omega \rightarrow \omega1 + \omega2$	Single solution with specific potential energy will produce two new spate solutions with different potential energy for each by select some of main steps of match result from original solution and other steps will be selected randomly, so single solution will be divided to produce two single solution with j
On wall effective collision $\omega \rightarrow \omega'$	Single solution will be combined with other random solution where each solution has its own potential energy to produce a new with different potential energy from original solution.

The proposed algorithm with CRO consists of a number of steps. Each of these steps has its fitness function value to find best matching for face recognition.

First step is initialization for face recognition using CRO. Through first step, a random selection is used to select pivot point, excluded features array and areas weight. Then the fitness function value is evaluated for each solution. As mentioned before, after each iteration of CRO execution, the population size will be different. We used 'parent size' variable to express population size. After each iteration it refers to generation size which produce new different solutions with different matching value.

The kinetic energy value must be defined in this stage, and molecule value also which must be between 0 and 1. Pseudo code for Initialization phase is given in Figure 6.

```
1 //initialization phase
2 Set Max values for different variable, C[i][j]: maximum
capacity
3 parentSize, iterationNumber, max value for variables
4 HIT= 0
5 β = parentSize/2
6 α = parentSize/2
7 KE = parentSize/1.5
8 Generate molecule ∈ [0, 1]
9 parentGenerating(C[i][j], parentSize ).
```

Figure 6: Pseudo Code for Initialization phase

Second step of face recognition using CRO refers to iteration stage. The main goal here is to generate different solutions based on initial solution. The initial values are produced through initialization phase, and new generated solution must be compared with all other solutions. Some of these new solutions will be better than original solution. The potential energy is objective function or fitness function for generated solution. Through this stage, different molecules are selected based on the value of variable β, where the value of β will be selected randomly between "0" and "1". As shown in Figure 7 for pseudo-code at line number 1, if the value of the variable β is larger than the value of the variable "molecule," then one molecule will be selected. Otherwise, two molecules will be selected; each of them is a separate solution. The pseudo-code for second phase of Face recognition using CRO algorithm is given in Figure 7.

```
for (int i=1 to iterationNumber) // Iteration phase
1 Generate b ∈ [0, 1]
2 if b > Molecule then
3 Randomly select one parent
4 if (HIT > α) then
5 Decomposition ()
6 else
7 OnWallIneffectiveCollision ()
8 end if
9 else
10 Randomly select two molecules
11 if (KE<=β && parentSize >=2) then
12 Synthesis ()
13 else if (parentSize >=2)
14 IntermolecularIneffectiveCollision ()
15 end if
16 end if
17 HIT++
18 KE—
```

Figure 7: Pseudo code for second phase of Face recognition using CRO.

Third phase of face recognition using CRO refers to reaction phase, which is the final stage. When reaching to a specific value

and to number of predefined iterations, then this algorithm is stopped. Next, different solutions must be found and each of them has its potential energy which is number of matching faces. The output from this stage is the best solution, with highest objective function or matching value. This solution contains specific value for all variables: a pivot value, excluding feature points and weight values for all face areas. Number of selected solutions will lead to different interactions between molecules or solutions. When number of selected molecules is one, then the possibility of some interactions that depends on single molecule interaction will be as follows:

On-wall-infective collision: This interaction depends mainly on single molecule interaction with the wall of container which contains all different solutions. The output from this interaction is a new molecule with different structure produced ω → ω' as in [17]. Pseudo-code for on-wall effective collision through our proposed idea is shown in Figure 8.

```
1 for (j=1 to number of cluster /2)
2 Fi'=generate new Solution randomly
3 end for-loop
4 //Compute objective function for the new molecule
5 PE' = objective Function(Fi')
6 // the new solution or dismiss
7 if (PE'> PE) then
8 destroy Fi
9 return Fi'
10 else
11 dismiss Fi'
```

Figure 8: Pseudo-code for on-wall effective collision

Second type of interaction refers to decomposition interaction. It deals with second interaction, which depends on single molecule with molecule containers. The difference here is that the output from this interaction is to divide the original molecule into two different molecules with same structure as follow: **ω → ω1 + ω2.** The pseudo-code for this interaction is shown in Figure 9.

If there are more than one molecule selected, the possibility for other interactions to occur between different selected molecules happens. These interactions between molecules are Synthesis interactions which happens when two molecules of solutions in our scheme interact with each other to generate new single solution from interacted solutions as follow: **ω1 + ω2 → ω'.**

3. Experimental Results

Both proposed ideas to enhance matching results for face recognition technique were implemented using java-programming language on Net Beans 8.1, through our implementation. We compared matching values of original face recognition algorithm with our proposed scheme. There are different results based on different parameter. First result deals with face recognition using Genetic algorithm for different iterations with same generation size. Obtained results collected from implementing this approach were compared with results from original scheme of face recognition using same data sets as shown in Table 3.

```
Function: Decomposition
Input: One Molecule/ One Solution
Output/ Two new molecules – two new solutions
1 // generate the first solution F1'
2 Copy the original solution Fi into F1
3 for (j=1 to Cluster No/2)
4 F1'= Randomly generate Clusters
5 end for-loop
6 // objective function for the new molecule
7 PE1'=ObjectiveFunction(F1)
8 if (PE1'> PEi ) then // solution confirmed
9 destroy Fi
10 return F1'
11 Else
12 dismiss F1'
13 end if
14 // generate the second solution F2'
15 Copy the original solution Fi into F2
16 for (j= number of clusters/2 to number of cluster)
17 F2'= Randomly generate new flow
18 end for-loop
19 // objective function for the new molecule
20 PE2'=objectiveFunction(F2)
21 //confirm the new solution or dismiss
22 if (PE2'> PEi ) then // solution confirmed
23 destroy Fi
24 return F2'
25 dismiss F2'
```

Figure 9: The pseudo-code for this interaction

Pseudo-code for synthesis interaction is shown in Figure 10.

```
Function: synthesis ()
    Input: F1, F2, PE1, PE2 // two molecules
Output: F3: solution has greater value of PE
1 If (PE1>PE2) then
2 F3 = F1
3 destroy F2
4 else
5 F3 = F2
6 destroy F1
7 return F3
```

Figure 10: Pseudo-code for synthesis interaction

Table 3: Results for implementing Face recognition using Genetic algorithm for different number of iterations

Iteration Number	Matching Percent values Using Genetic	Original Face Recognition
3	77.50%	71.30%
6	82.50%	67.30%
9	85.60%	67.30%
12	87.10%	67.30%
15	88%	67.30%
18	91.20%	67.30%
21	93.20%	67.40%
24	95.10%	67.30%
27	95.80%	67.30%
30	95.30%	67.30%

Table 3 contains results for implementing Face recognition using genetic algorithm for different number of iterations from 3 to 30 iterations. Each experiment has a specific iteration number with **fix** generation size. Best match results achieved when number of iteration is equal to 30 iterations. As the number of iterations increased the accuracy matching result increase. Figure 11 presents the relation matching accuracy versus number of iterations for face recognition.

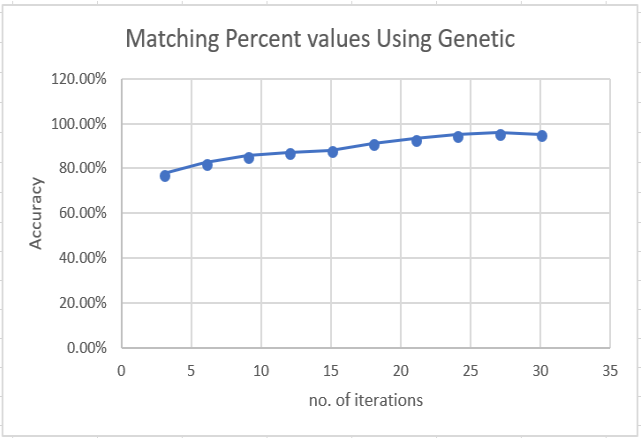


Figure 11: Relation between matching accuracy versus number of iterations for face recognition using Genetic algorithm

Table 4. Result for implementing face recognition using genetic algorithm for different generation size.

Generation Size	Matching Percent values Using Genetic	Original Face Recognition
5	82.83%	67.30%
10	83.74%	67.30%
15	86.80%	67.30%
20	89.85%	67.30%
25	89.6%	67.30%
30	89.9%	67.30%
35	91.3%	67.40%
40	92.1%	67.30%
45	92.6%	67.30%
50	93.7%	67.30%

Second experimental implementation for our proposed scheme refers to use different generation sizes with fixed number of iterations; generation size here starts from 5 to 50 where iteration size is the input for meta heuristic algorithm to generate new solutions. The results are shown below. Table 5 results refers to implementing other proposed schemes for enhancing face recognition results using chemical reaction optimization (CRO) for different number of iterations with fixed generation size.

Table 5: Results for implementing face recognition using CRO algorithm for different number of iterations

Iteration Number	Matching Percent value Using CRO	Original Face Recognition
3	71.65%	67.30%
6	76.71%	67.30%
9	83.30%	67.30%
12	85.86%	67.30%
15	82.1%	67.30%
18	87.86%	67.30%
21	87.1%	67.40%
24	89.8%	67.30%
27	91.92%	67.30%
30	92.83%	67.30%

Figure 12 presents the relation between number of iterations used for face recognition and matching value for Face recognition using CRO algorithm as follow.

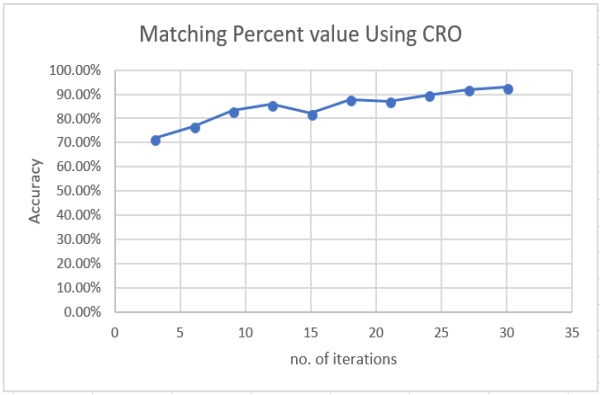


Figure 12: Relation between matching accuracy and number of iterations for our proposed scheme of face recognition using CRO algorithm.

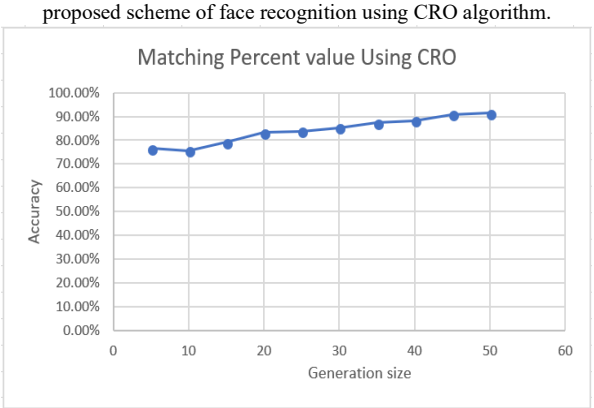


Figure 13: Relation between generation size and matching value for face recognition using CRO algorithm

Figure 13 presents the relation between matching percentage and generation size for face recognition using CRO algorithm. Other experimental implementation was for different generation sizes started from 5 to 50 with fixed value for iteration. Results for implementing face recognition using CRO algorithm are shown in Table 6.

Table 6: Results for implementing face recognition using CRO algorithm for different generation sizes.

Generation size	Matching Percent value Using CRO	Original Face Recognition
5	76.5%	67.30%
10	75.7%	67.30%
15	79.2%	67.30%
20	83.25%	67.30%
25	83.75%	67.30%
30	85.3%	67.30%
35	87.40%	67.40%
40	88.31%	67.30%
45	90.90%	67.30%
50	91.4%	67.30%

3.1. Comparison between CRO and Genetic algorithms for Face recognition

In order to compare between the accuracy of face recognition technique when using Genetic and CRO, the generated results are compared based on two criteria; number of iteration and generation size. The results for the number of iterations comparison are listed in table 7 and shown in Figure 14. It shows that in both algorithms, as the number of iterations increases, the recognition accuracy increases where CRO algorithm shows higher accuracy rate for the maximum number of iterations than Genetic algorithm.

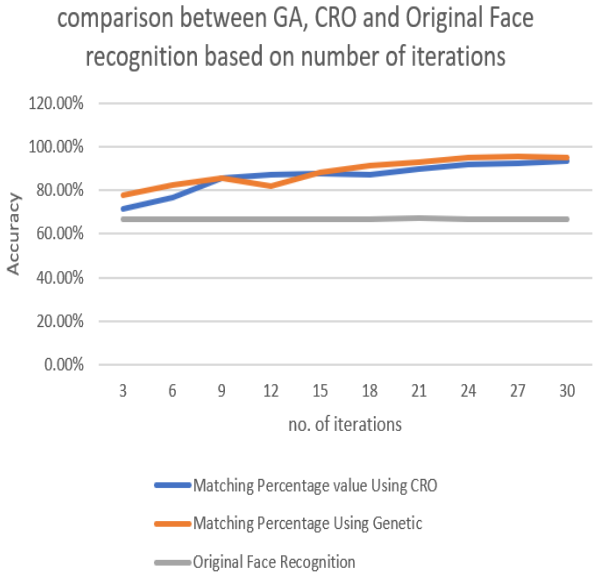


Figure 14: Comparative results of matching accuracy vs. number of iterations of both new approaches with original standard approach

Table 7: Comparison result for both approaches with original approach based on number of iterations

Iteration Number	Matching Percentage value Using CRO	Matching Percentage Using Genetic	Original Face Recognition
3	71.65%	77.5 %	67.30%
6	76.71%	82.5%	67.30%
9	85.86%	85.6%	67.30%
12	87.1%	82.1%	67.30%
15	87.86%	88%	67.30%
18	87.1%	91.2%	67.30%
21	89.8%	93.2%	67.40%
24	91.92%	95.1%	67.30%
27	92.3%	95.8%	67.30%
30	93.7%	95.3%	67.30%

Table 8: Accuracy of matching in percentage between both approaches based on Generation Size.

Generation size	Results for Proposed Approach using CRO	Results for Proposed Approach using Genetic	Original Face Recognition
5	76.5%	82.83%	67.30%
10	75.7%	83.7%	67.30%
15	79.2%	86.8%	67.30%
20	83.7%	89.6%	67.30%
25	83.7%	89.85%	67.30%
30	85.3%	89.9%	67.30%
35	87.4%	91.3%	67.40%
40	88.31%	92.1%	67.30%
45	90.90%	92.6%	67.30%
50	91.4%	93.7%	67.30%

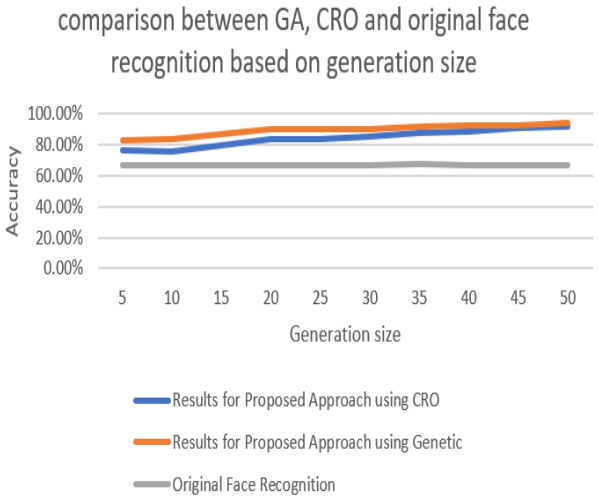


Figure 15: Results for Comparing the three approaches based on generation size.

Comparing the two algorithms for different generation size are listed in table 8 and shown in Figure 15. The comparison results

show that for bigger generation size the accuracy rate increases and CRO algorithm gives higher accuracy rate for the maximum generation size.

4. Results Discussion

The proposed work gives a new contribution in the face recognition technique which depend on three main criteria:

1. Selecting a pivot point.
2. Excluding some features into an exclude array base on the nature of the image.
3. Dividing the face into areas and assign a weight for each one.

Two heuristic algorithms were used in order to select best values for these variables, the Genetic and CRO. The aim of using such algorithm is to enhance searching results and get optimal solution. In our contribution, applying the heuristic algorithm on the face recognition technique gives more accurate results, as the algorithm searches for the best node in the image to be a pivot while the original technique select randomly a node as a pivot. The pivot node selected by the genetic and CRO is the node that is at the center of the image where we can evaluate distance between it and all other point in the image which will give an accurate estimation for the feature and thus will affect on the matching result and reduce error rate.

From the extracted features we can decide which features are important and which are not. For example, in the case where the image gas a special character like glasses the area around the eye will gives less matching results when it compared with the same image but without such glasses, so excluding this feature from the comparison will gives better result as shown previously.

By dividing the face image into a number of areas and assign a weight value for each one depending on the features for each area give a better accuracy for matching, some areas may be affected by different factors like brightness, shadow, etc. by assigning a low weight for those areas and higher value weight for the clear one affect much on the matching and increases accuracy.

The heuristic algorithms help on achieving that, as each algorithm depends on the number of iteration and generation size, the results show that as the number of iterations increase the matching accuracy increase as each algorithm training become near to its optimal solution. On the other hand, as the generation size increase the accuracy also increases, which means using more features and points from the image will increase matching results.

5. Conclusions

This paper presents a new methodology for face recognition by using two meta-heuristic algorithms: Genetic Algorithm (GA) and Chemical Reaction Optimization Algorithm (CRO). The aim of the proposed work is to enhance matching results for face recognition by increasing accuracy and decreasing error rate. Three criteria were taken into consideration throughout our work: the selection of the pivot point, excluding unnecessary features, dividing the face image into areas and assigning a weight value to each area. The GA and CRO were used to implement these criteria and generate a best solution. A dataset of 371 images was used for the training and testing phases. The results for each run were compared with the original standard face recognition technique and showed that GA and CRO enhance matching face recognition by achieving more than 20% accuracy over original technique.

References

[1] Aoun, N.B.; Mejdoub, M.; Amar, C.B. Graph-based approach for human action recognition using spatio-temporal features. J. Vis. Commun. Image Represent. 2014, 25, 329–338.

[2] El'Arbi, M.; Amar, C.B.; Nicolas, H. Video watermarking based on neural networks. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006; pp. 1577–1580.

[3] El'Arbi, M.; Koubaa, M.; Charfeddine, M.; Amar, C.B. A dynamic video watermarking algorithm in fast motion areas in the wavelet domain. Multimed. Tools Appl. 2011, 55, 579–600.

[4] Wali, A.; Aoun, N.B.; Karray, H.; Amar, C.B.; Alimi, A.M. A new system for event detection from video surveillance sequences. In Advanced Concepts for Intelligent Vision Systems, Proceedings of the 12th International Conference, ACIVS 2010, Sydney, Australia, 13–16 December 2010; Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6475, pp. 110–120.

[5] Koubaa,M.;Elarbi,M.;Amar,C.B.;Nicolas,H.Collusion,MPEG4compression andframedroppingresistant video watermarking. Multimed. Tools Appl. 2012, 56, 281–301.

[6] M. S. Obaidat and N. Boudriga," Security of e-Systems and Computer Networks, Cambridge University Press, 2007.

[7] Mejda Chihaoui *, Akram Elkefi, Wajdi Bellil and Chokri Ben Amar, "A Survey of 2D Face Recognition Techniques", REGIM: Research Groups on Intelligent Machines, University of Sfax, National School of Engineers (ENIS), Sfax 3038, Tunisia; Elkfi@gmail.com (A.E.); wajdi.bellil@ieee.org (W.B.); chokri.benamar@ieee.org (C.B.A.) * Correspondence: mejda.chihaoui@ieee.org; Tel.: +216-5460-1073

[8] D. Yi, Z. Lei, and S. Z. Li, "Towards pose robust face recognition," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013, pp. 3539–3545.

[9] N. Jindal and V. Kumar, "Enhanced face recognition algorithm using pca with artificial neural networks," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 6, 2013.

[10] Pratibha Sukhija, Sunny Behal, Pritpal Singh, "Face Recognition System Using Genetic Algorithm", International Conference on Computational Modeling and Security (CMS 2016).

[11] Ravi Subban, Dattatreya Mankame, Muthukumar Subramanyam," Genetic Algorithm based Human Face Recognition", Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC

[12] https://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/xm2vts/xm2vts_markup.html.

[13] D.E. Goldberg, Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley, Reading, MA, 1989.

[14] John H. Holland 'Genetic Algorithms', Scientific American Journal, July 1992.

[15] KalyanmoyDeb, 'An Introduction To Genetic Algorithms', Sadhana, Vol. 24 Parts 4 And 5.

[16] A.Y.S. Lam, V.O.K. Li, "Chemical reaction optimization: a tutorial", Memetic Computing 4, 2012, pp. 3–17.

[17] Albert Y. S. Lam and Victor O. K. Li, Chemical reaction optimization: a tutorial, Memetic Computing, Vol. 4, No. 1, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 116-127 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

FPGA Implementation of Ultra-High Speed and Configurable Architecture of Direct/Inverse Discrete Wavelet Packet Transform Using Shared Parallel FIR Filters

Mouhamad Chehaitly[1,*], Mohamed Tabaa[2], Fabrice Monteiro[1], Juliana Srour[3], Abbas Dandache[1]

[1]*Université de Lorraine, LGIPM, Metz, France*

[2]*Multidisciplinary Research and Innovation Laboratory (LPRI), EMSI Casablanca, Morocco*

[3]*Lebanese University, Beirut, Lebanon*

ARTICLE INFO

Article history:
Received: 10 July, 2018
Accepted: 08 September, 2018
Online: 25 September, 2018

Keywords:
Mallat binary tree algorithm
DWPT
IDWPT
FIR filter
Parallel-Pipeline Architecture
VHDL-RTL modeling
FPGA

ABSTRACT

This work presents new pipeline-parallel, generic and configurable parallel hardware architectures for the Direct/Inverse Wavelet Packet Transform (DWPT/IDWPT) independent of any specific family of wavelets, implemented in FPGA technology using a parallel architecture of direct FIR filter. We propose in the following paper, new P-parallel structures for the DWPT and IDWPT transforms based on the Mallat binary tree algorithm. Therewith, we developed a P-parallel/modified direct FIR filter architecture using pipeline-parallelize and hardware resource sharing, which provides not only ultra-high speed data processing but also a limited amount of hardware as resources are shared between filters and the bidirectional wavelet packet transformation. This model follows two important strategies: I) a powerful structure pipeline/P-parallel using strict data management and interleaving, II) sharing hardware at different levels in the transformation and between the two DWPT/IDWPT transformations. These architectures are modeled in VHDL at RTL modeling level. They are generic and fully configurable: at synthesis and post-synthesis.

The simulation results show an acceleration of data processing to an approximate value of P-Parallel multiplied by frequency with lower used resources. Furthermore, the impact of tree depth and filters order on throughput is very light due to the linearize architecture of our model. The synthesis was achieved using the Intel Quartus Prime Lite Edition software and targeting the Intel Altera Cyclone FPGA technology.

1. Introduction

1.1. Sate of art

Wavelet transform are considered the rightful heir to Fourier transform, since they are the best signal analysis tool in time and frequency domain.

From a historical point of view, the founder of the wavelets world was J. Morlet in 1980 who developed a new transform called "Wavelet Transform (WT)" to solve the limitations of the different versions of the Fourier transform on signal analysis passage from the time to the frequency domains [1, 2]. The numerical application of the wavelet transform is born with the work of Morlet and Grossman [3], while the idea of wavelet seems to originate in the work of Gabor and Neumann in the late 1940s. Mallat [1, 4], Meyer [5], Chui [6], Daubechies [7] and others, contributed the pioneering work reported in the early monographs.

The discrete version of the wavelet transform called "Discrete Wavelet Transform (DWT)" mixes both the notions of time and frequency scaling using a compact kernel (bit wave = Wavelet). Mallat [1] generalized the discrete wavelet transform that introduce more flexibility on the time scale of data analysis and lead to the concept of Discrete Wavelet Packet Transform (DWPT).

Currently, the architecture of DWPT and Inverse DWPT (IDWPT) are unsuitable in most fields like image processing, video processing, noise reducing, encryption, coding, modulation and others. This explains the large number of researches and

* Mouhamad Chehaitly, Email: che.liban.tly@hotmail.com

restrains the usage of DWPT in large applications domains, which leads us to develop a new hardware architecture to improve the performance of hardware implementation of DWPT/IDWPT. The major challenges for modern systems can be summarized as: (i) high throughput rates (increasing the volumes of data processed, especially in multimedia application) and (ii) low cost hardware (limited amount of hardware resource) while still providing high performance.

1.2. Related works

Abundant researches can be found in the literature about the hardware implementations of wavelet packet transform. The first work related to the hardware implementation of DWT was recorded in 1994 [8] Denk and Parhi described an orthonormal DWT architecture, which used a lattice Quadrature Mirror Filters (QMF) structure and digit-serial processing techniques. Similarly, Hatem et al. [9] proposed a parallel/sequential QMF architecture in VLSI for the DWT, to reduce the overall numbers of multipliers. After that, the research evolution was started.

Since DWPT retains all the advantages of the discrete wavelet transform, it has becomes an essential standard tool for signal, image and video processing. For this reason, the DWPT topic has attracted a lot of attention in research area. Notably, the most implementations of DWPT are based on the concept of FIR (Finite Impulse Response filter) filter banks [10]. In the literature, many works are addressing the implementation of DWPT/IDWPT in hardware devices such as Processors, GPUs (Graphic Processing Units) or FPGAs, to take full advantage of the powerful and flexibility offered by the DWPT. In this context, the year 1999 viewed the first work to implement the DWPT on a processor [11] but in the last recent years, FPGA became a popular target technology for many works, which become doubtlessly the target implementation technology for DWPT and IDWPT. However, the implementation of DWPT/IDWPT (based of FIR filter banks) has many caveats, which make it hard to fulfill a high throughput rate and low area consumption simultaneously.

To fulfill these challenges, many works implement the DWPT or IDWPT transformations on FPGA like the work of Wu and Hu [12], the authors implemented DWPT/IDWPT using Embedded Instruction Codes (EIC) and strict number of multipliers and adders for the symmetric filters (two multipliers and four adders). In [13], IDWPT architecture was presented based on classical recursive pyramid algorithm (RPA) by using lifting transformation and polyphase decomposition. In [14], a fast and configurable DWPT was presented based on FIFO inputs, dual-port memory, multipliers and adders.

An inspirational architecture of DWPT was presented in [15] by implement the tree algorithm, their implementation on FPGA was discussed in [16]. In [17], the same authors based on two modes developed pipelined architectures of the direct and inverse DWPT transform: serial-word mode and parallel-word mode. In [18], Farahani and Eshgho presented word-serial pipeline architecture of DWPT/IDWPT transforms by using parallel filter structure. To speed up the transform, the same authors updated their architectures by using high-pass and low-pass filters. In addition, they implemented the wavelet packet transform on FPGA by using three types of multipliers [20].

To increase the performance, in [21] a VLSI architecture of DWPT was presented which based on frame-partition architecture. In [22], authors implemented a flexible architecture of DWPT and IDWPT based on register interface and a multiplexing structure.

An algebraic integer architecture of Daubechies DWPT was presented in [23], this architecture was developed to compute one and two dimensions of Daubechies wavelet order 6 (Db6) with adder-less.

1.3. Contributions

This rich literature around DWPT and IDWPT reveals the importance of wavelet transforms in the current and future applications, which motivates us to deeply excavate in DWPT/IDWPT research area. Furthermore, most of recent approaches present some advantages in terms of flexibility, scalability, reliability, or low costing but not all together within the same architecture.

However, our aim is to develop new architecture generic and configurable (in synthesis and post-synthesis) of DWPT and IDWPT, which can provide an ultra-high speed data processing with minimum hardware usage. We already proposed our first pipeline and configurable architecture of DWPT in 2015 [24] and of IDWPT in 2016 [25] based on Mallat binary tree scheme [4] without any presence of parallelization tools. We transformed successfully an exponential algorithm of DWPT/IDWPT, based on classic Mallat binary tree, to a linear algorithm with the conservation of sample speed processing but with decreasing the usage of hardware by a smart managing and interleaving of data along the depth of transformations.

To increase the throughput, we propose in this literature massive pipeline-parallel architectures of DWPT (P-DWPT architecture) and of IDWPT (P-IDWPT architecture). Moreover, a massive pipeline-parallel DWPT-IDWPT architecture with massive sharing of resources: the core of the architecture is based on sharing hardware resources between P-DWPT and P-IDWPT transforms, and also based on a smart sharing of computational resources (multipliers and adders) between the approximation/details functions related to filter banks.

Those architectures are generic and fully configurable relative to the tree depth, the filters order, the quantization, and selection way of processing method (P-DWPT or P-IDWPT).

To evaluate the performance of our architectures, we study the effect of four parameters: “degree of parallelization, transformation depth of DWPT or/and IDWPT, order of low-pass and high-pass filters, and order of quantization” on the “operating frequency or throughput rate” (computational performance) and “resource consumption” (used hardware or hardware cost). Furthermore, we compared our synthesis results with the existing in literature.

1.4. Paper organization

The rest of this literature is organized as follows: Section II, presents a brief overview of wavelet packet transform concepts and the main challenge of parallelization in our context. Section III is dedicated to present our proposed pipeline-parallel architectures and related design paradigms. Then, we summarize the results

obtained for these architectures on FPGA implementation. At the end, we finalize with the comparison, conclusion and the perspective works.

2. Background

From mathematical point of view, wavelet packet transform decomposes the signal into a two sub-signals: detailed signal and approximated signal. To achieve the transforms there are several methods, the famous one is the multi-resolution or multi-cadence analysis algorithm proposed by Mallat [1], which called "Mallat binary tree" or "pyramid algorithm" of wavelet packet transform. This multi-resolution analysis is based on cascade of low-pass and high-pass digital FIR filters. By definition, using Wavelet Transform (WT) theory is that each signal $x(t)$ can be presented by projecting it into a series of scaled and translated functions $\Psi_{s,\tau}(t)$. The original single function is called "Mother Wavelet" and all scaled and translated functions are obtained by translation and expansion of the mother function, as presented in equation (1):

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \quad (1)$$

where $s \in R^*, \tau \in R$.

The value of s affects strictly the bandwidth of Mother Wavelet. When s is less than 1 ($s < 1$), wavelet variance decreases and the basis function Ψ contracts (used to analyze the high frequency signal). While for $s \geq 1$, wavelet variance increases and the basis function Ψ stretched (used to analyze the low frequency signal).

The Continuous Wavelet Transform (CWT) of $x(t)$ is defined as:

$$(s,\tau) = f, \Psi_{s,\tau} > \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} x(t)\Psi^*\left(\frac{t-\tau}{s}\right)dt \quad (2)$$

where * denotes a complex conjugation.

From the wavelet theory, the CWT is too redundant that make it impractical. To solve this problem, in 1992 Daubechies [26] create discrete wavelets transform (DWT) from the CWT by discretizing of scalable and translatable variables as shown in equation (3):

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}}\Psi\left(\frac{t-ks_0^j\tau_0}{s_0^j}\right) \quad (3)$$

Where the number of wavelets coefficients is finite. To discretize the scalable and translatable variables, we use the dyadic sampling concept that mean we usually choose $s = 2s_0, s_0 = constant$ and $\tau_0 = 1$ (give us a dyadic sampling in time). The Discretize of the translation and dilation contraction parameters of the wavelet in (2) leads to DWT presented in (4):

$$(j,k) = 2^{-j/2}\int_{-\infty}^{+\infty} x(t)\Psi^*(2^{-j}t-k)\,dt \quad (4)$$

where $j, k \epsilon Z$.

Another important function called "scaling function" φ is similar to the wavelet function $\Psi(t)$. According to the multiresolution wavelet theory, the φ is decomposed into finer and finer detail (in multiresolution stairs). As seen in equation (5), the $\varphi(t)$ has also two integer subscripts or parameters j and k.

$$\varphi_{j,k}(t) = 2^{j/2}\,\varphi(2^j t - k) \quad (5)$$

where j specifies the magnitude $2^{j/2}$as well as the scale $2^j t$ of the function, and k specifies the position (integer location, translation or shift) of the function.

In fact, DWT in signal analysis theory can be implemented with a different ways that is presented in equation (4) by using the concept of non-uniform filter bank. The filtering operation of input signal $x(t)$ is done iteratively and generates two separates sub-signals with two different spectrum, the upper half of the spectrum contains the high frequency component of input signal which is analyzed by the wavelet function and the lower part contains low frequency component of the same input signal which is continued to the next stage. The corresponding obtained coefficients that present the first filtered signal are called detailed coefficients and the second filtered signal are called smooth coefficients. Classically, both low-pass digital filter h and high-pass digital filter g are obtained from the scaling function and their corresponding mother wavelets.

We suppose h and g like a FIR filters non-recursive with L length, the transfer functions of h and g can be represented as follows:

$$H(Z) = H_0 + H_1 z^{-1} + H_2 z^{-2} + \cdots + H_{L-1} z^{-(L-1)} \quad (6)$$
$$G(Z) = G_0 + G_1 z^{-1} + G_2 z^{-2} + \cdots + G_{L-1} z^{-(L-1)} \quad (7)$$

Where z^{-1} denotes a delays of $1 \times sampling\ period$.

The orthogonal multiresolution decomposition of DWT can be carried out efficiently by using Mallat tree algorithm or pyramid algorithm [1]. This structure constitutes a bank of filters in QMF analysis, where the scaling functions and wavelet functions are realized using scale relations (6) and (7) and can be given as:

$$\phi(t) = \sum_n h(n)\phi(2t-n) \quad (8)$$
$$\Psi(t) = \sum_n g(n)\Psi(2t-n) \quad (9)$$

Where $\phi(t) = [\phi_1(t)\phi_2(t) \ldots \phi_r(t)]^T$ and

$\Psi(t) = [\Psi_1(t)\Psi_2(t) \ldots \Psi_r(t)]^T$ forms the set of scaling functions and their corresponding wavelets. The suffix r denotes the number of wavelets and is dubbed as multiplicity.

According to the wavelet theory based of filter bank concept, any arbitrary signal $x(t)$ can be expanded into a sum of scaling and wavelet functions. The discrete wavelet transform of target signal $x(t) \in L^2(\mathbb{R})$ is given by:

$$(t) = \sum \sum_{(i,l)\in l} \sum_{k\in Z} d_l^i(k)\Psi_{l,k}^i(t) \quad (10)$$

2.1. DWPT under focus

The decomposition DWPT at each resolution level, based on Mallat binary tree, can be presented as tree shape. To demonstrate a general transform of DWPT based of Filter Bank concept, we use Figure 1.

The input signal ($D_0^0[k]$) in coming data level will be decomposed into a high frequency signal (smooth or

approximation coefficient $D_1^0\ [k]$) and a low frequency signal (detailed coefficient $D_1^1\ [k]$) in level 1 by using low-pass / high-pass filters respectively. Then data path will be down sampling by a factor of two (lead to half size of original signal) and so on for the other levels (number of level equal the depth in Mallat tree algorithm). As a global view, the corresponding wavelet coefficients in different level are derived as follows:

$$D_l^{2i}(k) = \sum_n h(n) D_{l-1}^i(2k-n) \quad (11)$$

$$D_l^{2i+1}(k) = \sum_n g(n) D_{l-1}^i(2k-n) \quad (12)$$

where $i = 0, 1, \ldots, (2^{(l-1)} - 1)$ and $h(n)$ and $g(n)$ are low pass and high pass filters, respectively.

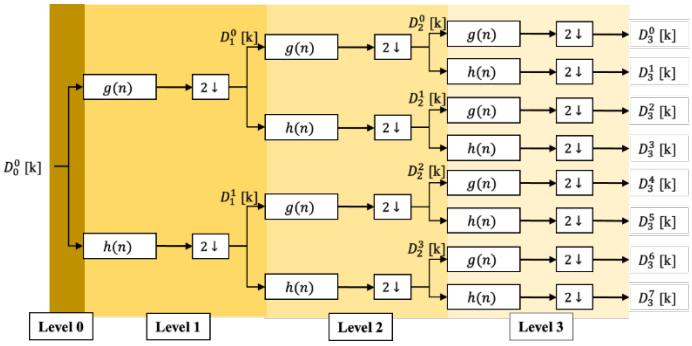


Figure 1. Three-level DWPT decomposition tree

Considering the FIR (Finite Impulse Response) non-recursive implementation scheme of length L for the h and g filters, the corresponding transfer functions can be represented as equations (6) and (7). The number of filters coefficients depends on the mother wavelet. The choice of this later depends upon the required application and its properties. Examples are Daubechies family where the filter length is $2N$ (Order N is strict positive integer), and Coiflets family where the filter length is $6N$ (Order N is strict positive integer), etc.

From the wavelet theory, the reconstruction of original signal that decomposed by direct DWPT is achieved by using the inverse of wavelet packet transform (IDWPT). Like the direct way, the reconstruction operation is also performed by using an iterative method. This mean, for each pair coefficients at level $l+1$ of the tree we can calculate the wavelet packets coefficients at the previous level l as shown in equation 13:

$$D_l^i(k) = \sum_n \bar{h}(n) D_{l+1}^{2i}(2k-n) + \sum_n \bar{g}(n) D_{l+1}^{2i}(2k-n) \quad (13)$$

In sample representation of equation (13), it can be represented in Figure 2.

For example to reconstruct the D_{j+1}^0coefficient at level $j+1$, we used the approximation D_j^0 and detailed D_j^1 coefficients at the previous level j. The reconstruction operation is provided by adding zeros and convolving the results with the reconstruction filters.

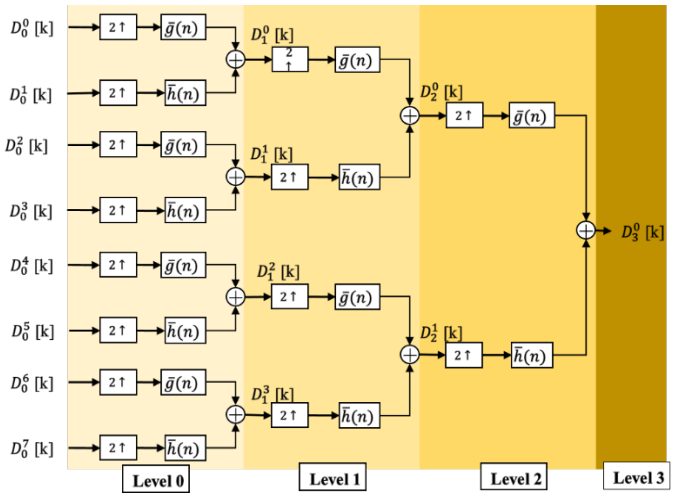


Figure 2. Three-level IDWPT reconstruction tree

To perfectly reconstruct the original signal (before the DWPT transformation), it is sufficient to use the concept of QMF filters that satisfy the following relations:

$$g(i) = (-1)^i h(L-1+i)$$

$$\bar{h}(i) = h(L-1-i)$$

$$\bar{g}(i) = (-1)^{i+1} h(i) = g(L-1+i) \quad (14)$$

2.3. Parallelization challenge of DWPT and IDWPT

In order to achieve our aim to develop a high performance hardware architecture of DWPT and IDWPT, we think for the first time by a classic way to parallelize the Mallat binary tree as shown in figures 3 and 4.

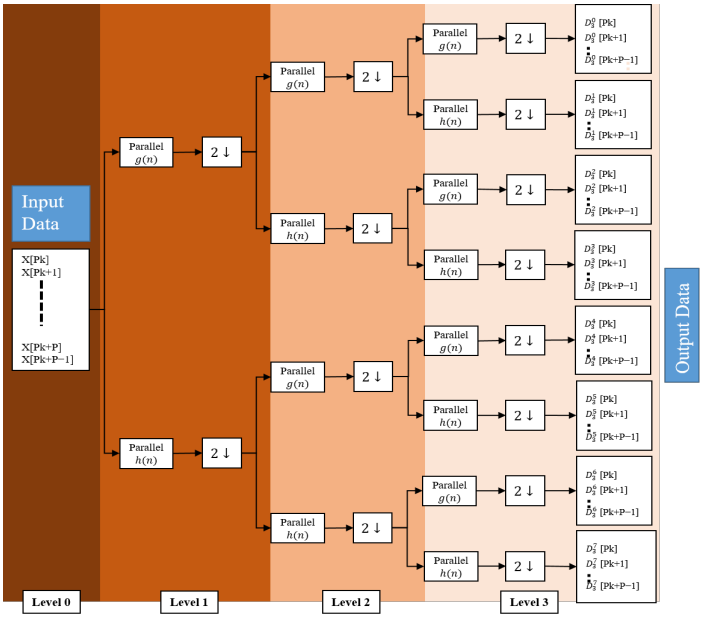


Figure 3. Classic P-parallel, Three-level DWPT

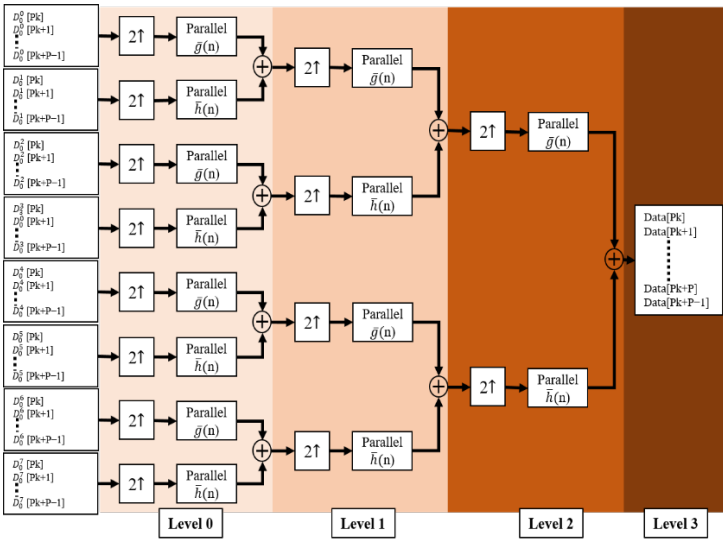


Figure 4. Classic P-parallel, Three-level IDWPT

A traditional possible parallelization solution is presented in figures 3 and 4, where we show a three-level decomposition, reconstruction tree respectively, with theoretical P-parallel degree and a P-parallel filters bank consists of wavelet functions to be able to treat P sampling in each slot time. This classic P-parallel architecture is unrealizable because the number of reconstruction filters and/or decomposition filters increases exponentially as a function of depth order. In precision, the number of filters needed to implement this architecture is around $P * (2^{depth+1} - 1)$. For example, with degree of parallelism $P = 16$ and $depth = 5$, we need 1008 low/high pass filters for DWPT and another 1008 low/high pass filters for IDWPT. The implementation of 2016 filters is a potential problem, which make it unable to implement.

3. Our Parallel-Pipeline architectures with sharing resources

- At a given stage of the Mallat binary tree, the data rate processed by any filter is twice as fast as that of any filter on the adjacent stage on the input side in Figure 1.
- While, the data rate processed by any filter is twice as slow as that of any filter on the adjacent output-side stage in Figure 2 (factor 2 upper sampling from stage to stage).
- In addition, the amount of data processed in a k-level is $2^{depth} - k$ times the amount to process at level 1. Thus, the total amount of data to be processed in level k is the same as in the first or last level ($2^{-k} \times 2^{k} = 1$).

Based on this big regularity that provide high throughput rate with lower hardware resources, we build an evolution parallel-pipeline architecture of DWPT and IDWPT based on Mallat binary tree as shown in section III. To achieve our objective that ensure high throughput with low hardware consumption in our parallel architecture of DWPT and IDWPT, we have to modify the high/low pass FIR filters.

Under the strategy to reduce area consumption, instead of using two filters we merge in the same architecture the functionality of low-pass and high-pass filters in a single block filter. Furthermore, we modify the single block to process P sampling in two-clock cycle. In figure 5, we present our P-Parallel modified transposed FIR filter.

This new architecture is linearized to be similar to coding theory where the Serial transposed FIR filter is like a Single-Input Single-Output (SISO) system and the Parallel transposed FIR filter is like a Multiple-Input Multiple-Output (MIMO) system. Therefore, this heterogeneous architecture provides the processing of P inputs signals and consequently P outputs signals, in each clock cycle.

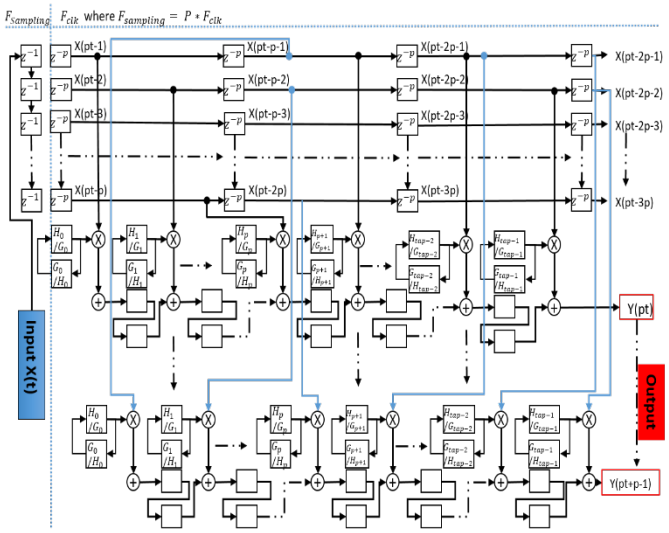


Figure 5. Architecture of a single modified FIR filter

The main difference with the filters, proposed in [24, 25] is related to the handling of the low-pass and high-pass filter coefficients. Furthermore, this filter requires to be feed by correctly scheduled data by a smart shift of data between different stages to serve P sampling in each clock cycle.

Respecting the z^{-p} buffer order, it is the most sensitive and the core of the serial FIR filter to parallel operation. This role of manage and interleaving data is devoted to the key block in our model. We called this block "buffers block" situated between the filters in different level. The structure of a single buffer is dependent on the direct or inverse DWPT.

In our following proposed architecture, we respect these constraints:

- The degree of parallelism and the characteristics of the target card (technology) must be taken into consideration in order not to deposit available resources.
- The degree of parallelism must respect the dyadic rule, that is, $P = 2^{x}, \forall\, x \in \mathbb{N}^{+}$. Where P affects directly the data managing between different levels of the transformation tree and simplifies the up/down sampling operation.
- To synthesis our proposed hardware architecture, we used

Altera Quartus premium lite edition software that is targeted on Altera FPGA belonging to the Cyclone V family with a speed grade of –7.

3.1. First proposed architecture: Parallel-Pipeline architecture of DWPT with sharing resources

This part is dedicated to present a P-Parallel DWPT architecture, which provides high throughput, by using the modified transposed P-parallel digital FIR filters (presented in figure 6).

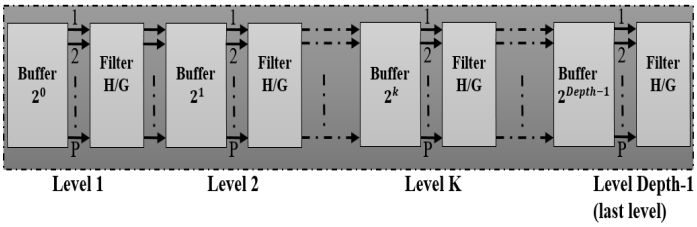


Figure 6. Datapath diagram of the proposed P-parallel DWPT architecture

As we mention below, this modified filter or H/G block, presented in Figure 5, ensures the processing of the same amount of data on any stage k in the original Mallat binary tree.

Concerning our plan to reduce the hardware usage, the functionality of high-pass filter and low-pass filter is provided by the H/G blocks. Consequently instead of using two separate similar filters (high-pass filter and low-pass filter), we propose using a single filter in an alternatively process. Hence, this alternatively process (on consecutive clock cycle) works by taking a sample for the $H(Z)$ and then for $G(Z)$ and so on. The critical point in this model is to manage correctly the data between filters and different levels, this role is devoted to a specific buffers situated between the filters in different stages. The structure of blocks buffer is shown in figure 7.

A. Buffer Block structure of DWPT

The k parameter in figure 7 is related to the number of the stage in which the buffer is implemented. Then in each level k in figure 6, the buffer blocks is built up on two types of shift register (2^{k-1} positions shift for each one): a “fast shift buffer” which takes data from the previous stage, and a “slow shift buffer” which feeds its data from its own stage filter. Overall, the buffer size depends on the parallel degree P and the level k on which it is implemented.

In this new scheme, the fast buffer (and hence in the slow buffer too) are stored, by interleaving ordered disposition, filtered data of all related stage. On each 2^k cycles, only half samples can be transferred from the fast buffer part to the slow buffer part. Furthermore, each sample flows out from the slow buffer to the next stage must be presented twice (granted by the low frequency rate of the slow buffer) in order to be processed by both $H(z)$ and $G(z)$ filters. To manage different control signal and the data interleaved in different stage, we developed a control unit called "$Control_block$".

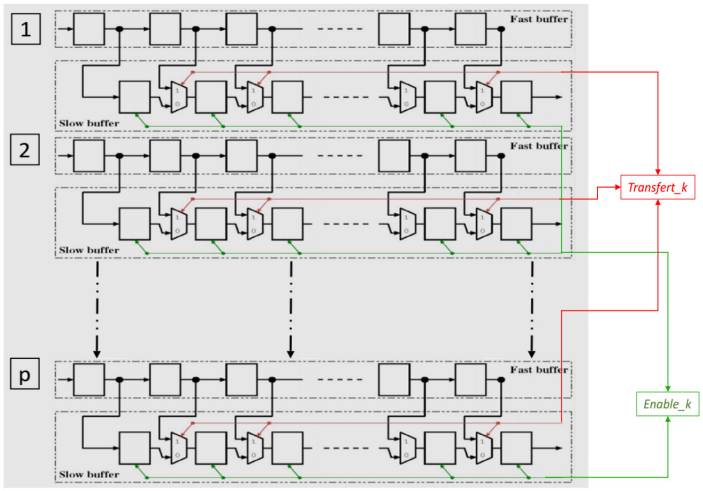


Figure 7. Structure of the DWPT buffer block in stage k

The "$Control_block$" is dedicated to generate and manage the various controls signals "$enable_k$" and "$transfer_k$" related to stage K.

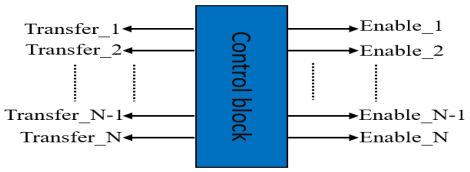


Figure 8. Control unit block

B. Synthesis results

As we mentioned, this architecture is fully configurable in synthesis by three parameters: the wavelet scale (the tree depth), the order of the filters and the filter coefficient quantization (generic parameters in the VHDL-RTL model). Moreover, it is partially configurable after synthesis of filters coefficient values corresponding to different wavelet family (with the same filter order). This means we can change the type of wavelet without re-synthesis our architecture (the coefficients are loaded dynamically after synthesis) where it was not the case for all the previous work.

In tables 1, 2 and 3, the configuration parameters are presented as 3-tuples: Depth of DWPT tree, Order of filter, and Quantization

with the synthesis results of implementation for 4, 8 and 16 parallel DWPT architecture. Although the highest rate of our architecture is not requiring any memory or DSP blocks.

Table 1. Implementation results of 4-parallel DWPT architecture

Design parameters (Depth, Order *Quantization*)	**Data processing rate**	**Resources usage (l_e, l_r)**
(2, 2, 5)	*415.36*	*(471,296)*
(3, 2, 5)	*400.86*	*(756,510)*
(4, 2, 5)	*389.56*	*(1204,899)*
(2, 4, 5)	*403.52*	*(879,456)*
(3, 4, 5)	*342.46*	*(1299,719)*
(4, 4, 5)	*375.24*	*(1941,1171)*
(2, 16, 5)	*357.99*	*(3299,1416)*
(3, 16, 5)	*369.83*	*(4794,1924)*
(4, 16, 5)	*3425.69*	*(6397,2614)*
(2, 2, 16)	*208.14*	*(2571,905)*
(3, 2,16)	*198.62*	*(4216,1599)*
(4, 2,16)	*209.89*	*(5850,2853)*
(2, 4,16)	*201.18*	*(5038,1324)*
(3, 4,16)	*194.60*	*(7521,2260)*
(4, 4,16)	*184.120*	*(10374,3636)*
(2, 16,16)	*180.41*	*(4902,4402)*
(3, 16,16)	*220.31*	*(6805,5729)*
(4, 16,16)	*213.63*	*(9107,7752)*

Table 2. Implementation results of 8-parallel DWPT architecture

Design parameters (Depth, Order *Quantization*)	**Data processing rate**	**Resources usage (l_e, l_r)**
(2, 2, 5)	*583.03*	*(1109,504)*
(3, 2, 5)	*557.23*	*(1699,935)*
(4, 2, 5)	*562.25*	*(2754,1531)*
(2, 4, 5)	*586.20*	*(2120,897)*
(3, 4, 5)	*506.77*	*(3050,1197)*
(4, 4, 5)	*526.91*	*(4603,2023)*
(2, 16, 5)	*499.48*	*(7689,2447)*
(3, 16, 5)	*504.56*	*(12176,3166)*
(4, 16, 5)	*478.26*	*(14956,4571)*
(2, 2, 16)	*282.35*	*(6079,1696)*
(3, 2,16)	*294.41*	*(9279,2735)*
(4, 2,16)	*300.60*	*(13489,5011)*
(2, 4,16)	*289.86*	*(12032,2582)*
(3, 4,16)	*273.15*	*(17549,3965)*
(4, 4,16)	*238.45*	*(24311,6363)*
(2, 16,16)	*302.62*	*(11263,7856)*
(3, 16,16)	*325.85*	*(14750,11451)*
(4, 16,16)	*309.31*	*(21314,13091)*

3.2. Second proposed architecture: Parallel-Pipeline architecture of IDWPT with sharing resources

A new P-parallel architecture of IDWPT will proposed in the following part, which can provide ultra-high speed data processing and low cost resources consummation.

Table 3. Implementation results of 16-parallel DWPT architecture

Design parameters (Depth, Order *Quantization*)	Data processing rate	Resources usage (l_e, l_r)
(2, 2, 5)	*718.13*	*(3668, 652)*
(3, 2, 5)	*710.42*	*(6019, 960)*
(4, 2, 5)	*709.04*	*(8655, 1243)*
(2, 4, 5)	*536.55*	*(5991, 1689)*
(3, 4, 5)	*517.53*	*(8380, 2363)*
(4, 4, 5)	*516.33*	*(11181, 2601)*
(2, 16, 5)	*464.04*	*(30012, 4881)*
(3, 16, 5)	*454.15*	*(37172, 6575)*
(4, 16, 5)	*454.92*	*(38374, 7680)*
(2, 2, 16)	*455.49*	*(11116, 3395)*
(3, 2,16)	*447.30*	*(17679, 4330)*
(4, 2,16)	*446.25*	*(25389, 4830)*
(2, 4,16)	*336.75*	*(31336, 5137)*
(3, 4,16)	*331.99*	*(39361, 7764)*
(4, 4,16)	*341.59*	*(42687, 10342)*
(2, 16,16)	*302.46*	*(26408, 15572)*
(3, 16,16)	*303.44*	*(33859, 20959)*
(4, 16,16)	*298.79*	*(36348, 24456)*

The reconstruction process of the P-parallel IDWPT architecture is simply the reversed form of P-parallel DWPT. Considering the diagram in Figure 4, we can observe that any filter of level k is able to process $P * 2^{(depth-k)}$ times the amount of data to be processed on level 1. Thus, the total amount of data by all filters to be processed in level k is the same as in level 1 ($P * 2^{(depth-k)} * 2^{-(depth-k)} = P$). It is the same remark that is observed at the DWPT in the previous part (III.1 a). Moreover, we can notice that at any filter, in a given level, half the data is processed than its neighbors on the output side, and twice its neighbor on the input side. This implies that the construction of structure tree have repeated functionally blocks and the complexity of the filters is the same for all stages.

To eliminate the exponential evolution of filters number as function of depth order, we linearize and serialize the filtered data in different stages. Consequently, instead of using $P * 2^k/2$ low pass filters and $P * 2^k/2$ low pass filters to filter data in stage k, we just implement only one modified filter in each stage. As shown in figure 10, the number of used modified transposed FIR filter bank increases linearly a function of depth order.

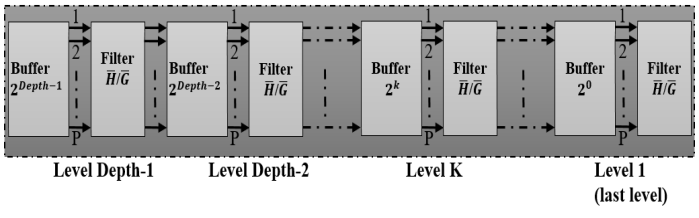


Figure 10. Data path diagram of the proposed P-parallel IDWPT architecture

The modified single filter ($\frac{\bar{H}}{\bar{G}}$ Filter Block) can process the same amount of data on stage k in the original tree (Figure 4). In figure 10, $\frac{\bar{H}}{\bar{G}}block$ is the high-pass filter and low-pass filter, which are

related to $h(n)$ and $g(n)$, in equation (16).

The structure for the modified FIR filter is the same as that presented in figure 5. The only difference is the coefficients of FIR filters concerning $\bar{H}(Z)$ and $\bar{G}(Z)$. This filter provides the same functionality of that presented by Mallat in their binary tree, we also reduce their occupied area by merging low-pass and high-pass reconstruction filters function in the structure in a single block. The processing of filtering is similar to that used in decomposition part (sub section III-1) where we take a sample for $\bar{H}(Z)$ filter and for $\bar{G}(Z)$ filter by an alternatively process (on consecutive cycles) and so on.

Consequently, this modified block can process one sample in two clock cycles. To ensure the best interleaving and managing of data in different stages, we developed a key block in our entire model that is the "Buffer Block", which is situated between the filters in different levels.

A. *Buffer Block structure of IDWPT*

The most important thing in our P-parallel IDWPT (Figure 10) is, as previously, the link between neighbor levels, which does not need any reorganize of data set. We use the same concept of buffer blocks as illustrated in Figure 6, so the size of the buffer shown in figure 11 depends on the stage in which it is implemented. The buffer block is then built up on two sub-blocks in each k positions (k is an indicator of the stage where the block buffer is implemented).

Then in each level k in figure 10, the buffer block is built up on two types of shift register (2^{k-1} positions shift for each one): a "fast shift buffer" takes data from the previous stage, and a "slow shift buffer" which feeds its data from its own stage filter. Overall, the buffer size depends on the parallel degree P and the stage k in which it is implemented.

The "fast shift buffer" is faster than "slow shift buffer" where it achieves P-shift on each clock cycle while the "slow shift buffer" achieves P-shift on two-clock cycles. To handle this latency, we used an enable signal called "$enable_k$" which manages the slower shift rate in level k. To control the transfer data from the fast buffer to slow buffer, we insert a counter signal 2^k to activate the $transfer_k$ signal on every 2^k cycles. This transfer operation is performed P times at each stage.

The synthesis results of implementation of our P-parallel IDWPT architecture based on modified P-parallel FIR filter structure will be the subject of this part.

This architecture of P-parallel IDWPT is fully configurable in synthesis by three parameters: the wavelet scale (the tree depth), the order of the filters and the filter coefficient quantization (generic parameters in the VHDL-RTL model). Moreover, it is partially configurable after synthesis of filters coefficients values corresponding to different wavelet families (which have the same filter order). The resources consumed by our proposed architecture resources are given as 2-tuples (l_e, l_r): where l_e stands for logic elements and l_r for logic registers. To generate the differences counter, the different control signals "$enable_k$" and "$transfer_k$" related to stage k, we proposed also like in figure 8 a block control unit.

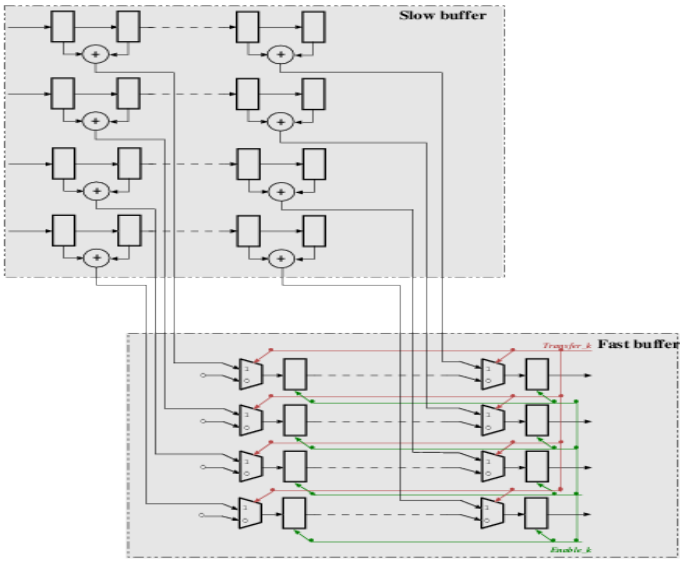


Figure 11. Structure of the IDWPT buffer block in stage k with degree of parallelization P=4

Table 4. Implementation results of 4-parallel IDWPT architecture

Design parameters (Depth, Order *Quantization*)	Clock frequency	Resources usage (l_e, l_r)
(2, 2, 5)	*205*	*(109, 186)*
(3, 2, 5)	*201.82*	*(166,312)*
(4, 2, 5)	*196.16*	*(244,505)*
(2, 4, 5)	*152.88*	*(265,286)*
(3, 4, 5)	*152.58*	*(379,442)*
(4, 4, 5)	*153.37*	*(483,665)*
(2, 16, 5)	*144.03*	*(1447,886)*
(3, 16, 5)	*137.44*	*(1983,1222)*
(4, 16, 5)	*136.24*	*(2457,1625)*
(2, 2, 16)	*132.36*	*(578,582)*
(3, 2,16)	*135.34*	*(833,972)*
(4, 2,16)	*133.69*	*(1102,1572)*
(2, 4,16)	*104.57*	*(1594,902)*
(3, 4,16)	*102.77*	*(2174,1388)*
(4, 4,16)	*100.61*	*(2772,2084)*
(2, 16,16)	*94.14*	*(7719,2822)*
(3, 16,16)	*92.08*	*(10557,3884)*
(4, 16,16)	*90.49*	*(13469,5156)*

B. *Synthesis results*

In tables 4, 5 and 6, the configuration parameters are presented as 3-tuples: Depth of DWPT tree, Order of filter, and Quantization with the synthesis results of implementation for 4, 8 and 16 parallel DWPT architecture. Although the highest rate of our architecture does not require memory or DSP blocks.

3.1. *Third proposed architecture: Parallel-Pipeline architecture of IDWPT-DWPT with sharing resources and shared FIR modified filter*

This part is dedicated to present another version of parallel DWPT/IDWPT architecture. In some applications and especially in digital communication, we find that there is a need to implement

the DWPT and IDWPT transforms on the same FPGA split. Certainly, we can implement our architecture (P-parallel DWPT and P-parallel IDWPT) on the same board to ensure the transformation functionality using the same constraints of work with ultra-high throughput and low resources consumption (consequently low power consumption). We notice a big regularity in the architecture of P-parallel DWPT (Figure 6) and P-parallel IDWPT (Figure 10), particularly in the modified FIR filters on the different levels where the structure of modified FIR filter is the same and the difference is particularly in the loaded coefficients after synthesis. We think to develop, at last but not at least, new architecture, which retains all features from previous architectures (ultra-high throughput, generic, configurable with depth and with order of filters and with quantization). In addition, the same architecture can be implemented with P-parallel DWPT or P-parallel DWPT by the selection of the active way as shown in figure 12.

Table 5. Implementation results of 8-parallel IDWPT architecture

Design parameters (Depth, Order Quantization)	Clock frequency	Resources usage (l_e, l_r)
(2, 2, 5)	*207.04*	*(202, 347)*
(3, 2, 5)	*195.77*	*(301,574)*
(4, 2, 5)	*198.73*	*(409,863)*
(2, 4, 5)	*147.15*	*(527,527)*
(3, 4, 5)	*148.39*	*(705,804)*
(4, 4, 5)	*147.65*	*(813,1123)*
(2, 16, 5)	*136.44*	*(2955,1607)*
(3, 16, 5)	*133.05*	*(3802,2184)*
(4, 16, 5)	*131.56*	*(4288,2683)*
(2, 2, 16)	*128.75*	*(1140,1095)*
(3, 2,16)	*123.08*	*(1529,1806)*
(4, 2,16)	*128.04*	*(1828,2711)*
(2, 4,16)	*99.98*	*(3098,1671)*
(3, 4,16)	*98.87*	*(4094,2542)*
(4, 4,16)	*98.95*	*(4733,3543)*
(2, 16,16)	*90.16*	*(15416,5127)*
(3, 16,16)	*86.02*	*(20279,6958)*
(4, 16,16)	*86.1*	*(23194,8535)*

Table 6. Implementation results of 16-parallel IDWPT architecture

Design parameters (Depth, Order *Quantization*)	Clock frequency	Resources usage (l_e, l_r)
(2, 2, 5)	*197.47*	*(389,668)*
(3, 2, 5)	*195.35*	*(573,1096)*
(4, 2, 5)	*194.97*	*(742,1576)*
(2, 4, 5)	*147.54*	*(1091,1008)*
(3, 4, 5)	*142.31*	*(1410,1526)*
(4, 4, 5)	*141.98*	*(1552,2036)*
(2, 16, 5)	*127.6*	*(5894,3048)*
(3, 16, 5)	*124.88*	*(7300,4106)*
(4, 16, 5)	*125.09*	*(7536,4796)*
(2, 2, 16)	*125.25*	*(2183,2120)*
(3, 2,16)	*123.02*	*(2704,3472)*
(4, 2,16)	*122.71*	*(3016,4986)*
(2, 4,16)	*92.6*	*(6154,3208)*
(3, 4,16)	*91.29*	*(7730,4848)*
(4, 4,16)	*93.93*	*(8383,6458)*
(2, 16,16)	*83.17*	*(30619,9736)*
(3, 16,16)	*83.44*	*(39257,13104)*
(4, 16,16)	*82.16*	*(42143,15290)*

In this scheme, we represent a new pipeline-parallel architecture of DWPT/IDWPT with shared P-parallel FIR filter. Furthermore, instead of the implementation of 2*Depth P-parallel FIR filter in this architecture, we used the half of hardware architecture with the conservation of the ultra-throughput rate.

The buffers block are similarly presented in previous part in P-parallel DWPT and P-parallel IDWPT. Respectively, Mux_Data and DeMux_Data blocks are implemented to manage data between the buffers and the shared P-parallel FIR filter in P-parallel DWPT or P-parallel IDWPT. The Mux_Filter blocks are multiplexer, used to precise the transform direction and to manage the loading FIR filters coefficients after synthesis.

A. Synthesis results

To evaluate the performance of our P parallel DWPT/IDWPT architecture, we just implement the Blocks of P-parallel FIR filters, which gives us a percentage of the resources quantity consumption that have been dispensed by using a shared P-parallel FIR filters.

The compilation results of area consumption for different values of the configuration parameters, that mean tree depth, filter order, coefficient quantization (in number of bits), and parallel degree are presented in table 7 with different parallel degree.

Table 7. Implementation results of 4, 8 and 16-parallel degree of Parallel FIR architecture

Design parameters (Depth, Order *Quantization*)	FIR Resources usage (l_e, l_r)		
	P=4	**P=8**	**P=16**
(2, 2, 5)	*(41,27)*	*(128,64)*	*(298,106)*
(3, 2, 5)	*(41,28)*	*(138,67)*	*(315,110)*
(4, 2, 5)	*(49,30)*	*(138,70)*	*(315,113)*
(2, 4, 5)	*(53,37)*	*(148,83)*	*(371,127)*
(3, 4, 5)	*(53,38)*	*(167,86)*	*(371,131)*
(4, 4, 5)	*(61,40)*	*(167,89)*	*(371,135)*
(2, 16, 5)	*(112,94)*	*(281,189)*	*(701,319)*
(3, 16, 5)	*(112,96)*	*(298,195)*	*(701,322)*
(4, 16, 5)	*(119,98)*	*(307,195)*	*(701,327)*
(2, 2, 16)	*(198,83)*	*(302,107)*	*(368,161)*
(3, 2,16)	*(198,88)*	*(328,112)*	*(368,166)*
(4, 2,16)	*(218,93)*	*(353,117)*	*(368,172)*
(2, 4,16)	*(257,114)*	*(378,138)*	*(431,337)*
(3, 4,16)	*(257,119)*	*(428,143)*	*(431,198)*
(4, 4,16)	*(297,124)*	*(428,148)*	*(431,203)*
(2, 16,16)	*(557,278)*	*(760,314)*	*(1164,408)*
(3, 16,16)	*(557,283)*	*(806,319)*	*(1164,408)*
(4, 16,16)	*(593,288)*	*(806,324)*	*(1164,413)*

From table 7, we notice that the resources consumption are the same with the same order filter (which is expected). Moreover, with half usage of P-parallel FIR filter in P-parallel DWPT/IDWPT, we gain resources percentage from 3 to 5 % for a depth = 4 and filter order = 16, which is significant especially with the large depth and high filter order (like discrete Meyer wavelet).

This is a very important result because classically it is too difficult to implement in the same board simultaneously both parallel architectures.

4. Comparison

To evaluate the performance of each architecture, we make a comparison chart (Table 8) with several DWPT and IDWPT architectures designs achieved in literature.

A quantitative and comprehensive comparative are presented in Table 8, where we can observe that:

- Our architectures are implemented without using any memory or DSP block where most of other architecture required that.
- Our models DWPT/IDWPT architectures are generic and fully configurable. They are suitable for all wavelet families (where the wavelet families are simplified by the filters orders) where it is not the case for all previous works.
- In our architectures; the quantization, the depth order and the filters order are theoretically unlimited (depends on the limit of manufacturing technology)

From these observations, table 8 presents the superiority of our architectures that are clearly presented on the high bite-rate driven by high frequency (around 200MHz).

5. Conclusion

In this work, we have proposed three powerful P-parallel/pipeline configurable architecture of Direct Discrete Wavelet Packet Transform (DWPT) and Inverse Discrete Wavelet Packet Transform (IDWPT) based on Mallat binary tree using bank filter concept, which provide high throughput with minimal hardware resources. The considered problem is to accelerate data processing in clock cycle and decrease the total hardware resources used.

To solve these problems, we develop a P-Parallel modified FIR filter based on transposed FIR filters that can share hardware resources between low-pass and high-pass filters (by merging their functionality). The effective data path maintains a short critical path allowing high operation frequency to be reached, where synthesis results indicate that our architecture provides a very high speed data processing with minimum resources. For example: for a parallel degree P=16, depth order = 2, filter order = 2 and order quantization = 5 we reach a very high bit rate equal to 3159.52 Mega samples. We also developed a new effective architecture for both DWPT and IDWPT implemented in the same programmable board, conserving the high throughput performance with more decreased hardware resources consumption.

Those proposed architectures (IDWPT and DWPT) are fully configurable at synthesis as function of P-parallel degree, depth (number of tree stages), filter order and filter coefficient quantization (generic parameters in the VHDL-RTL model). Furthermore, they are re-configurable after synthesis by loading the filter coefficients, which depend on different wavelet family during operations (after synthesis) that providing high flexibility of DWPT/IDWPT transforms.

Perspectives

In this work, we present a different parallel version of hardware implementation in FPGA of DWPT/IDWPT. This work is still in progress where we are constructing another generation of DWPT/IDWPT implementation, which can provide more performance and can be used in new application domains.

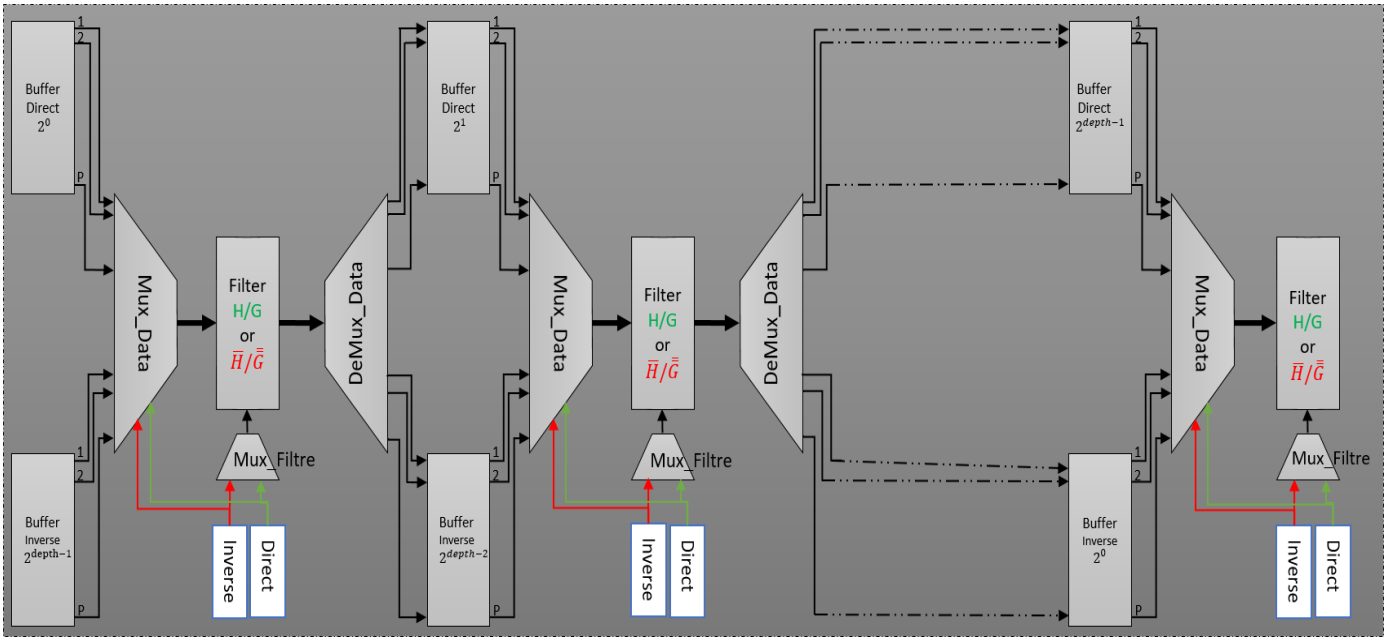


Figure 12. Proposed data-path diagram of both P-parallel DWPT and IDWPT architecture.

Table 8. Comparison of proposed architectures with existing DWPT and IDWPT architectures.

	Tze-Yun et al.[28]	Marino et al. [29]	Mohanty et al. [30]	Madishetty et al. [23]	Wang et al. [27]	Wu et al. [12]	Meihua et al. [21]	Proposed architecture N°1	Proposed architecture N°2	Proposed architecture N°3
Wavelet	Daub-4	Quadri-filter	Daub-4	Daub-6	Lifting-Db4	Quadri-filter	Quadri-filter	arbitrary	arbitrary	arbitrary
DWPT/ IDWPT	Both	DWPT	DWPT	DWPT	Both	Both	DWPT	DWPT	IDWPT	Both
Logic cell	N/A	N/A	426	1040	N/A	30192 (logic gate)	1835 (logic unit)	3668 (logic element)	389 (logic element)	
Technology	Xilinx XC2V4000	N/A	CMOS 90nm	Xilinx Virtex 6	CMOS 180 nm	0.35 µm	Altera EP20K200E	Altera Cyclone IV	Altera Cyclone V	Altera Cyclone V
Max. Freq.(Mhz)/ Bitrate	N/A	N/A	20	306.15	20	100	29	Bitrate:718.13	Feq.: 197.47	~200
Memory	Yes	N/A	yes	yes	Yes	N/A	yes	No	No	No
Quantization (bits)	N/A	N/A	N/A	N/A		32	N/A	Test with 5 to 16 (up to unlimited theatrically)	Test with 5 to 16 (up to unlimited theatrically)	Test with 5 to 16 (up to unlimited theatrically)
DSP	N/A	N/A	N/A			N/A	N/A	No	No	No
Depth	3	N/A	N/A	4	3	3 (up to 6)	3	2, 3 and 4 (up to unlimited theatrically)	2, 3 and 4 (up to unlimited theatrically)	2, 3 and 4 (up to unlimited theatrically)
Parallel degree								4,8 and 16 (up to unlimited theatrically)	4,8 and 16 (up to unlimited theatrically)	4,8 and 16 (up to unlimited theatrically)

References

[1] S. Mallat, A wavelet tour of signal processing, Academic Press, 1999.

[2] R. Dilmaghani and M. Ghavami, "Comparison between wavelet-based and Fourier-based multicarrier UWB systems," IET Communications, Vol. 2, Issue 2, Feb. 2008. http://dx.doi.org/10.1049/iet-com:20070181

[3] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," SIAM journal on mathematical analysis, 15(4), 723-736, 1984. https://doi.org/10.1137/0515056

[4] S. Mallat, "Multiresolution approximations and wavelet orthonormal bases of L2(R)," Transactions of the American Mathematical Society, vol. 315, no. 1, pp. 69–87, 1989. https://doi.org/10.1090/S0002-9947-1989-1008470-5

[5] Y. Meyer, "Wavelets and operators," vol. 37 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1992.

[6] C.K. Chui, "An Introduction to Wavelets," Philadelphia, SIAM, vol. 38, 1992.

[7] I. Daubechies, Ten lectures on wavelets, Philadelphia, SIAM, vol. 61, pp. 198–202, 1992.

[8] T.C. Denk and K.K. Parhi, "Architectures for lattice structure based orthonormal discrete wavelet transforms," IEEE International Conference on Application Specific Array Processors, pp. 259-270, Aug. 1994. https://doi.org/10.1109/ASAP.1994.331798

[9] H. H. A. Hatem, M. El-Matbouly, N. Hamdy, K. A. Shehata, "VLSI architecture of QMF for DWT integrated system," Circuits and Systems, MWSCAS 2001. Proceedings of the 44th IEEE 2001 Midwest Symposium on (Volume: 2), pp. 560–563, 2001. https://doi.org/10.1109/MWSCAS.2001.986253

[10] G. Strang and T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, 1997.

[11] W. Xiaodong, L. Yongming and C. Hongyi, "Programmable wavelet packet transform processor," In Electronics Letters, vol. 35, no. 6, pp. 449-450, 18 Mar 1999. https://doi.org/10.1049/el:19990330

[12] B-F. Wu and Y-Q. Hu, "An efficient VLSI implementation of the discrete wavelet transform using embedded instruction codes for symmetric filters," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 9, pp. 936–943, September 2003. https://doi.org/10.1109/TCSVT.2003.816509

[13] G. Paya, M. M. Peiro, F. J. Ballester, V. Herrero and J. Cerda, "A new inverse architecture discrete wavelet packet transform architecture," IEEE, Signal Processing and Its Applications, 7803-7946, 443 - 446 vol.2, 2003. https://doi.org/10.1109/ISSPA.2003.1224909

[14] A. Jamin and P. Mähönen, "FPGA implementation of the wavelet packet transform for high speed communications," International Conference on Field Programmable Logic and Applications. Springer Berlin Heidelberg, 2002. https://doi.org/10.1007/3-540-46117-5_23

[15] M. Trenas, J. López, M. Sanchez, F. Argüello and E. L. Zapata, "Architecture for Wavelet Pack et Transform with Best Tree Searching," In Application-Specific Systems, Architectures, and Processors, 2000. Proceedings. IEEE International Conference on (pp. 289-298). IEEE, 2000. https://doi.org/10.1109/ASAP.2000.862399

[16] M. Trenas, J. López and E. Zapata, "FPGA Implementation of Wavelet Packet transform with Reconfigurable Tree Structure," Euro micro Conference, 2000. Proceedings of the 26th Volume 1, 5-7, pp. 244 - 251 vol.1, September 2000. https://doi.org/10.1109/EURMIC.2000.874639

[17] M. Trenas, J. López, E. Zapata and F. Argüello, "A Configurable Architecture for the Wavelet Packet Transform," Journal of VLSI signal processing systems for signal, image and video technology, Volume 32, Issue 3, pp 255-273, November 2002. https://doi.org/10.1023/A:1020221003822

[18] M. Farahani and M. Eshghi, "Architecture of a wavelet packet transform using parallel filters," TENCON 2006 - IEEE Region 10 Conference, p. 7-10, 2006. https://doi.org/10.1109/TENCON.2006.343965

[19] M. Farahani and M. Eshghi, "Implementing a new architecture of wavelet packet transform on FPGA," Proceedings of the 8th WSEAS International Conference on Acoustics & Music: Theory & Applications, Vancouver, Canada. 2007.

[20] M.A. Farahani, S. Mirzaei and H.A. Farahani, "Implementation of a reconfigurable architecture of discrete wavelet packet transform with three types of multipliers on FPGA," 24th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2011. https://doi.org/10.1109/CCECE.2011.6030704

[21] X. Mei-hua, C. Zhang-jin, R. Feng, and C. Yu-lan, "Architecture research and VLSI implementation for discrete wavelet packet transform," High Density Microsystem Design and Packaging and Component Failure Analysis, 2006. HDP'06. Conference on. IEEE, 2006. https://doi.org/10.1109/HDP.2006.1707554

[22] D. Garcia, M. Mansour and M. Ali, "A flexible hardware architecture for wavelet packet transform with arbitrary tree structure," IEEE Transactions on circuits and systems – II: Express briefs, vol. 60, no. 10, p. 657-661, October 2013. https://doi.org/10.1109/TCSII.2013.2277964

[23] S. K. Madishetty, A. Madanayake, R. J. Cintra and V. S. Dimitrov, "Precise VLSI Architecture for AI Based 1-D/ 2-D Daub-6 Wavelet Filter Banks with Low Adder-Count," IEEE Transactions on circuits and systems-I: regular paper, Vol. 61, No. 7, 1984 - 1993, July2014. https://doi.org/10.1109/TCSI.2014.2298283

[24] M. Chehaitly, M. Tabaa, F. Monteiro and A. Dandache, "A fast and configurable architecture for Discrete Wavelet Packet Transform," In Design of Circuits and Integrated Systems (DCIS), 2015 Conference on (pp. 1-6), 2015. https://doi.org/10.1109/DCIS.2015.7388599

[25] M. Chehaitly, M. Tabaa, F. Monteiro and A. Dandache, "A VHDL-RTL implementation for a fast and configurable design of inverse discrete wavelet packet transform," Design of Circuits and Integrated Systems (DCIS), 2016 Conference on, p. 1 – 5, 2016.

[26] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, "Image coding using wavelet transform," IEEE Transactions on image processing, 1(2), 205-220, 1992.

[27] C. Wang, J. Zhou,L. Liao, J. Lan, J. Luo, X. Liu and M. Je, "Near-threshold energy-and area-efficient reconfigurable DWPT/DWT processor for healthcare - monitoring Applications," IEEE Transactions on Circuits and Systems II: Express Briefs, 62(1), 70-74, 2015. https://doi.org/10.1109/TCSII.2014.2362791

[28] T.-Y. Sung, H.-C. Hsin, Y.-S. Shieh and C.-W. Yu, "Low-power multiplierless 2-D DWT and IDWT architectures using 4-tap Daubechies filters," In Proc. Seventh Int. Conf. PDCAT, pp. 185–190, 2006. https://doi.org/10.1109/PDCAT.2006.78

[29] F. Marino, "Two fast architectures for the direct 2-D discrete wavelet transform," IEEE Trans. Signal Process, vol. 49, no. 6, pp. 1248–1259, 2001. https://doi.org/10.1109/78.923307

[30] B. K. Mohanty and P. K. Meher, "Memory-efficient high-speed convolution-based generic structure for multilevel 2-D DWT," IEEE Trans. Circuits Syst. Video Technol., vol. 23, pp. 353–363, 2013. https://doi.org/10.1109/TCSVT.2012.2203745

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 128-139 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Developing A Conceptual Framework of Product-Service System Management Toward Firms' Sustainability for Indonesian Industrial Estate Firms

Christina Wirawan*,[1,2], Gatot Yudoko[1], Yuliani Dwi Lestari[1]

[1]*School of Business Department, Institut Teknologi Bandung, Bandung 40123, Indonesia*

[2]*Industrial Engineering Department, Universitas Kristen Maranatha, Bandung 40164, Indonesia*

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 14 September, 2018
Online: 25 September, 2018

Keywords:
Product Service System
Sustainability
Firm's Life Cycle
Indonesian industrial estate firms

ABSTRACT

Industrial estate firms have important roles in national development through the industrial development. In the course of their life, industrial estate firms ought to follow the firm life cycle pattern. The common pattern shows that the period, in which the firm experiences profit increase, will be directly followed by consistent profit and profit decrease. Should the strategy realization are not precisely implemented during each stage, especially the decreases period, it will be difficult for the firm to survive.

A strategy that can be implemented by industrial estate firms in order to be sustained is to manage Product-Service System (PSS). It is a well-known concept in creating opportunities for a new business model in which firms can offer bundles of products and services to their customers instead of merely tangible products. PSS can be distinguished into three types of categorization consist of different compositions of products and services and different business activities. Each type of PSS and business activities can be applied at the stages of the firm life cycle in accordance with its condition to maintain the company's sustainability at every stage of its life cycle. In this article, we construct a conceptual framework to help Indonesian industrial estate firm to be alert on the need to focuses on providing services instead of only selling products. This article also shows that different PSS management would be needed in particular stages of firm life cycle. Applying suitable PSS and business activities based on PSS in each stage of the firm life cycle can lead to gaining a competitive advantage and sustainable industrial estate firms. This article's novelty is to combine the PSS concept concerning each stage of firm life cycle at industrial estate firms.

1. Introduction

This article is an extension of our previous work published at 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) [1]. Our previous work discussed about types of PSS and business activities implementation in Indonesian industrial estate firms in order to help them sustain [1]. Industrial estates firstly initiated in Europe, such as England; Netherland; Germany; and the United States of America in the late 1800s [2]. They are built as a center and localization for industrial companies [2]. Following after, many countries also built industrial estates for several reasons such as environmental and economic growth [3]. Industrial estate firms have obligation to provide land; building; infrastructures; utility services [4]; and amenities [5] for industrial companies in the estate and community. Industrial estates, primarily built to support industrial growth [2] by spurring industrial growth [6] and efficient resources usage [5] while also aimed to protect the environment [3]. Based on their purposes, industrial estate firms becoming important for a nation and need to be sustained. However, along their lifespan, industrial estate firm will then follow the firm life cycle pattern [7].

Following the common firm life cycle, industrial estate firms will experience profit ups and downs [8]. At the beginning of the firm life cycle, industrial estate firms have land and building to be sold, thus they will be able to make enough profit. However, along the time period, if the business runs well, these land and buildings will eventually be sold out. Thus, in the later stages, industrial estate firms need to find other sources of profit to keep them

*Christina Wirawan, Ganesha 10, Bandung, Indonesia, +62 8122119907
Email: christina.wirawan@sbm-itb.ac.id

sustained. Many industrial estate firms have tried to sustain by expanding the industrial estate of building another one in a different location. But this strategy seems to be unsuitable for long-term business since the land will eventually run out and the industrial estate market will also be saturated. Moreover, the local government will also impose several regulations that will limit the building of industrial estate firms. Then, to help industrial estate firms to be sustained, the concept of PSS is seemed to be suitable.

PSS concept attempt to shift tangible products consumption into services with the same performance, function, and benefit [9]. By applying the concept of PSS, the production and consumption of tangible products can be reduced so as to reduce the environmental burden, providing opportunities for new business models [9,10], and improving the social conditions of the community through co-creation [11]. According to experts, PSS is divided into three types and for each type of there are several types of business activities. In its implementation, the company can choose the type of PSS and business activities. Each type and business activities have their own characteristic that will build different strategies. These strategies would be suitable with the condition of the industrial estate firms at each stage of the life cycle. The appropriate business activity at the beginning of the life cycle stage may be different from the appropriate business activity at the later life cycle stages. The concept of PSS that combines the tangible products with services is considered to be suitable for the firms to achieve sustainability in each stage of its firm life cycle. It can be achieved by combining the limited tangible products with services.

In Indonesia, the first industrial estate was built in Cilacap in 1968 and the number was continuously increasing each year. The rapidly increasing number of industrial estate firms is partly due to the fact that since 1989 the private sector has been allowed to build industrial estate firms in Indonesia. Until now there are about 232 industrial estates in Indonesia [12]. About 68.24% industrial estates were located in Java, the most populous and developed island in Indonesia, where Jakarta as Indonesian capital is also located [12]. Therefore the Indonesian government encourages further industrial estates development outside of Java [13]. This regulation is applied to increase the distribution of industrial, economic and welfare development, especially outside of Java [12].

In Indonesia, the importance of industrial estate roles is strengthened by the Constitution of Republic of Indonesia number 3 the year 2014 (*Undang-Undang Republik Indonesia*) on the industry, which states the development of the industry as one of the pillars of national economic development [14]. Indonesian industrial estates were built to support national industries. According to Indonesian government's regulations, industrial estates in Indonesia were built for several purposes, such as to accelerate the spread and equality of industrial development; to develop environmentally friendly industrial development; to encourage investment and industrial competitiveness; and to give certainty to the site [15]. With these regulations, it is obvious that industrial estate firms play important roles to support economic development [14] through raising companies efficiency; attract investors [16]; create job opportunities; foster innovation [3]; and support welfare enhancement in surrounding area. Industrial estates also help the government to control environmental impact and industrial development [15]. Considering the importance of the roles of industrial estate companies, it is necessary to think applicable efforts to support their sustainability.

The Indonesian government has already been aware of the important role of industrial estate firms by encouraging industrial estate firms' performance and the effort for sustainability. One of the efforts done by Indonesian government is by setting government regulation in 2009 and 2015 that remark sustainability in term of economic [14-15,17] and competitiveness; social [17]; community [6]; and environmental [15,17] which are deemed suitable with sustainable development concept [18]. As an effort to improve the performance and sustainability of industrial estates, Indonesian government through The Indonesian Industrial Ministry holding Indonesian Industrial Estate Award(s), which has been held at 2013 and 2015 [17].

Indonesian industrial estate firm will also face a life cycle pattern problem, which in a certain period they will experience a decrease in profit due to exhaustion or reduction in product sales. But given its important role, they need to be sustainable. As discussed previously, the PSS concept would be necessary to help industrial estate firms to be sustained.

There have been many studies and discussions about PSS related to the life cycle carried out by researchers. Most researchers write about PSS to prolong product and PSS life cycle and further to recycle or rebuilt them.PSS concept discussed managing PSS product life cycle stages: design, manufacturing, usage, remanufacturing or disposal [19-20] toward sustainability. While for of PSS life cycle, PSS management discussed for the beginning of life, middle of life and end of life [20-21]. Beginning of life stage consist of ideation, requirement, design, middle of life consist of realization, delivery, support, and end of life consist of evolution [20, 22].

However, the article that discussed PSS management to address the firm life cycle has not been found. Although commonly the pattern of product, PSS, and firm life cycle may be similar, the strategy to manage the PSS would be different due to the unique characteristic of each industrial estate firms. The article discusses the implementation of PSS at industrial estate firms has also could not be found. Therefore, this article will discuss PSS implementation at industrial estate firms.

As pointed earlier, this article is an extent of our previous work. In the previous work, we found that Indonesian industrial estate began to lead to the provision of services, in addition to merely selling products [1]. They already practice some of the business activities as an effort to gain new profit toward sustainability [1]. This article will extend the finding by considering the firm life cycle. Relation of the type of PSS and business activities practiced with the stages of the firm life cycle would be studied. This study is conducted due to the fact that for industrial estate firm, there are different characteristics in each stage of the firm life cycle that will logically differ on what are the suitable type of PSS that need to be implemented.

The objective of this article is to build a conceptual framework regarding sustainable issues of industrial estate firms in each stage of firm life cycle using the PSS concept. This article will study each PSS types and business activities implementation at 5 industrial estates firms to understand PSS implementation in

Indonesian industrial estate firms while testing the theoretical based framework. This article will contribute to help industrial estate firm by raising awareness and giving suggestion to the industrial estate firms so that they will be more focus on providing services as the alternative income generator. This article also highlights the differences of PSS concept that is needed among stages of firm life cycle to help industrial estate firms sustain in each stage. For the government, this article will give insight for industrial estate planning and policy-making to encourage sustainability of the industrial estate firms to suitably play important roles. For academic field, this article will contribute to combine PSS concept with life cycle concept in an effort to help industrial estate firms to sustain.

In the following section, the literature review will be discussed, especially about the concepts that will be used as the basis for the formation of the framework and the previous studies followed by the research objective and methodology. Furthermore, the conceptual framework of PSS management will be presented at each stage of product life cycle followed by a discussion on PSS management framework and practice in Indonesian industrial area. In the end, this article will be closed by the conclusion, including implications and further research directions.

2. Literature Review

As previously mentioned, in most countries, industrial estates managed by industrial estate firms, have been built to help companies to operate more efficiently due to they are providing facilities, amenities, and waste treatment infrastructure to support companies' in their operation [4,5]. Industrial estates are built with the objectives to encourage employment rate and economic growth [6,23], to attract new firms [5] and foreign direct investments [6], to foster technological learning and innovation [3], to spur an industrial dynamic [6], to efficient resource used, and to lower environmental burden [23]. With this supports, companies may benefit from the reduced cost of getting water, energy, material, etc., as well as disposal cost [24]. Another benefit that will be gained by the company is an opportunity to build relationships and make collaborations with other companies and stakeholders to build value co-creation [25].

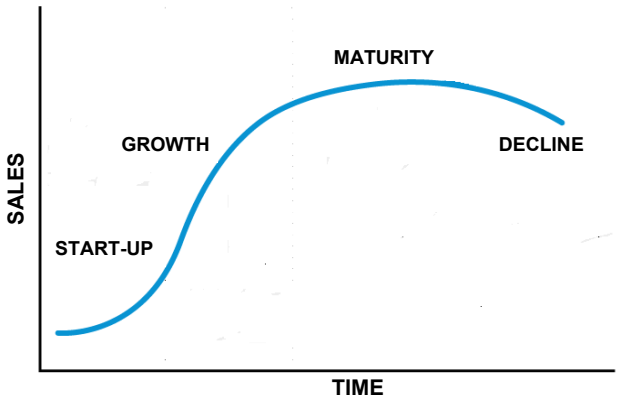


Figure 1. Common life cycle model[8], [26]

In the course of their growth and development, firms will experience the stages that form the general pattern, namely the firm life cycle [7]. Several authors identified divisions of firm life cycle stages. However, common stages are divided into start-up, growth, maturity and decline stages [26] with the S-shaped pattern [27]. The common pattern of the life cycle is as shown in Figure 1 [8,27]. At each stage of its life cycle, the company will experience different conditions of different pressures, threats, and opportunities [28]. Therefore, the company will need a different strategy and performance [29], structure [30], effectiveness criteria [31], profit-maximizing [27] to keep sustaining in each stage of the life cycle.

In Indonesia, industrial estate firms have an important role in national economic development. This is indicated by the Republic of Indonesian government regulation requiring industrial companies to operate within industrial zones administered by licensed industrial estate companies, except in limited circumstances [15]. While in Indonesian republican law, the industry is declared as one of the pillars of national economic development [14]. Thus, it is clear that the industrial estate firms have an important role to play in national economic development. Indonesian Ministry of Industrial and Trade at 1987 already stated that industrial estate firms have obligations to provide road network in industrial estate, drainage, installation of clean water supply along with distribution channels to each industrial plot, wastewater treatment plant which can accommodate all wastewater from factories in industrial estates, electricity supply installations and distribution networks, telecommunications networks, environmental impact control facilities, lighting, industrial estates corporate offices, fire-fighters, housing for industrial workers, social and public facilities in accordance with the relevant provisions of the relevant agencies [12]. The Indonesian government already paid attention to the importance of industrial estate firms, as well as efforts to support the Indonesian industrial estate firms' sustainability by issuing some regulations, such as that remark sustainability in term of economic [14–16] and competitiveness [17]; social [17]; community [6]; and environmental [15,17] which are deemed suitable with sustainable development concept [18], as well as formulating a roadmap for industrial estates and organizing industrial estate awards for high performer industrial estates [17].

The first industrial estate in Indonesia was built in Cilacap in 1968, and since then some new industrial estates are built each year and the number increases from year to year. Until now, prospects for industrial estates is still considered good since demand is rising [12]. Geographically, the industrial area in Indonesia is still not evenly distributed. Most of them or as much as 68.24% have been built on Java, while in Riau Island, Riau and Bintan is 17,17% [12]. In Sumatra, Kalimantan, Sulawesi, Ambon and Papua there are 7,73%, 3.43%, 2.58%, 0.43% and 0.43% respectively [12]. Java is attractive to industrial estate firms due to the density of the population thus able to provide advantages of localization and urbanization economy[32]. Moreover, the complete facilities and infrastructure in Java and the presence of the nation's capital making it more interesting for further gain on profit. With this condition, the Industrial Master Plan or RIPIN 2015-2035 prioritizes the development of industrial areas outside Java to encourage welfare distribution throughout Indonesia.

Indonesian industrial estate firm main business is selling products in the form of land and buildings that make most of the profits. In addition, industrial estate firms also provide a range of services to support and to assist companies in the industrial estates

such as renting warehouses, land, buildings, industrial vehicles and other industrial equipment; providing water supply, waste treatment, and disposal; shipping and delivery facilities, etc. [1]. Product in the form of land and building usually give enough profit for industrial estate firms, but it is still limited. Once the land and buildings are sold out, the profits will only come from the provision of services, which is, of course, lower than the sales of the products. Some industrial estate firms maintain their sustainability by expanding the land or opening industrial estates elsewhere. However, this strategy cannot be done continuously since in the near future, the land will run out. This condition is also barred by the limitation by the rules of fringe benefit that some allocated area is not supposed to be used for industries [33]. Therefore, a new business model that emphasizes the provision of services will be appropriate to be applied to industrial park companies in order to remain profitable enough to be sustained.

Related to the firms' life cycle concept, in the early stages of its life cycle, industrial estate firms will gain enough profit from product sales, but in the later stages where the product is sold out, the industrial estate company needs to find a new business model to earn enough profit to be sustained. Regarding the importance of the role of industrial estate companies in Indonesia in contributing to the development of the national economy, an effort to maintain their sustainability is highly needed. The concept of PSS that combines product sales and service provision will be suitable for application in industrial estate firms, especially in the later stage of the firm life cycle.

PSS is a system of products; services; network; and infrastructure designed to meet customer needs and cause lower environmental impact than traditional business models while seek to increase firm competitiveness [34]. PSS promote dematerialization [9,35]. It is the attempt to decrease tangible products consumption, and wherever possible change it into the bundle of products and services [9]. PSS also attempt to tangible products' useful life and increase the utilization of tangible products [10] that both will reduce tangible product usage. With PSS concept, consumer or user will receive relatively equal function, performance, and benefit as they buy tangible products in a different way. PSS concept attempt to decreased environmental burden and meeting the needs of the world's population that increased dramatically caused by the sharp increase in population and consumption of large quantities of products in recent decades [9].

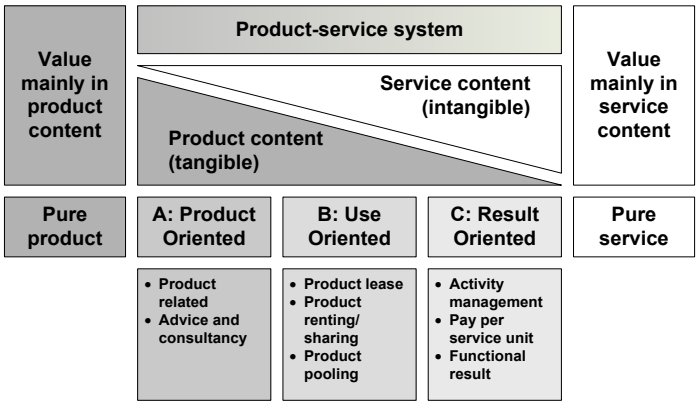


Figure 2. Types of Product-Service System (PSS) [35]

To attain their sustainability, PSS would be appropriate to implemented by industrial estate firms since it brings out an opportunity to find new added value and competitiveness. PSS can also increase customers' satisfaction by offering customized product and services, building a good customer relationship, and creating faster innovation opportunity [35]. All of these conditions can lead industrial estate firm to gain competitive advantage and then sustainability. A reference [36] stated that land and building selling in industrial estates need to be shifted into services in the future because of the availability of the limited products.

According to experts, the PSS is categorized into three different types, namely product-oriented, use-oriented, and result-oriented [10,36] as shown in Figure 2. Each of the type related to particular business activities and particular composition of products and services.

The first type is Product-oriented PSS. Product-oriented PSS is the type of PSS that most closely resembling a traditional business, where products ownership transferred to the customer as the main business, but with the addition of services to extend products' life and retrieve them when the product is unusable for recycling[10]. Reference [35] stated 2 business activities of product-oriented PSS, namely product-related services and also advice and consultation to prolong the product's useful life as an effort to decrease the product consumption [35]. Example of product-oriented PSS implementations in industries is Eastern Energy, which provides energy management; consumption and process monitoring; utility awareness; and training to help their customer maintain the products that will prolong their life cycle [10]. The other example is Polyplank AB that creates cheap, recyclable and moisture-resistant composite material to be used for paper mill's core plug from plastic waste [37]. Polyplank AB also takes back the core plug to be reused four times before need to be replaced [37]. With this system, products consumption can be reduced.

The second type is use-oriented PSS. The Use-oriented PSS is a PSS type in which the tangible product is used, but the ownership of the product is not transferred to the customer for the purpose of improving the utilization and extending the useful life of the product [10]. Business activities in this type are product lease, product renting/ sharing, and product pooling [35]. Examples of this type are vehicle sharing group, where vehicle provided by the provider and used by several customers by sharing scheme [10]. In Seoul, a public bicycle system operated to reduce energy consumption, raise community health status as well as reduce bicycle used and budget for buying bicycle by individuals [38]. The other example is Conway's water purifier which built a rental system with the regular visiting maintenance service for water purifier [39]. With this rental system, the customer does not need to buy the product, and other customers can use the product when he/she does not intend to use it. Regular visiting of maintenance will help the product to be in good condition and can be used for a longer time.

The last type is result-oriented PSS. Result-oriented PSS is the type of PSS with the greatest service composition. In this type, customers buy services in the form of capabilities and results [10]. Business activities in this type are activity management, pay-per-service unit, and functional result [35]. Example of this type o PSS implementations is Xerox pay-per-copy system, where the

customer will lease the copy machine and they will pay per copy they made counted by the machine [10]. Xerox will also provide maintenance service and will take the machine back after the leased in finished. With more or less similar system, Electrolux also create a system, which customers pay initial fee for placing the washing machine in their place, then payment would be made per wash [10]. Electrolux also is responsible for the maintenance, reparation, and another financial service of the washing machine [10]. With this system, the washing machine useful life will be prolonged, since the cost for maintenance and repair done by the experts, while customers benefit from efficient cost.

Among these three types of PSS, there would be differences in product and service proportion, as well as business activities. Product-oriented PSS has the largest proportion of the product, followed by the use-oriented PSS, thus conversely, result-oriented PSS has the largest service proportion. Business activities of each type of PSS would be appropriate to be applied at each stage of the firm life cycle of industrial estate firms to help industrial estate firms to be sustained in each stage of the life cycle.

This article will discuss the conceptual framework of PSS management at each stage of the firm life cycle. The discussion will take the context of Indonesian industrial estate firms that have important roles in Indonesian economic development as stated in the Indonesian Constitution Number 3 the year 2014. This article will provide a novelty on the application of PSS types at the stages of the firm life cycle within the context of industrial estate firms in Indonesia.

3. Research Objectives and Methodology

From the beginning, the Government of Indonesia built industrial estate firms with the main objective of controlling industrial development, controlling the environmental burden, overcoming the limitations of industrial support, and accommodating housing needs around industrial sites [36]. Later on, the purpose of an industrial estate building is growing primarily to develop the industry [36] in an attempt to support national economic growth [14]. Different from most countries, starting from 1989 until now, industrial estates in Indonesia are allowed to be built by the private sector [36]. With this permission, industrial estates in Indonesia are developing even faster. The fast development is due to capital addition and new innovation from private firms. In future development, industrial estates in Indonesia are directed and expected to adopt the concept of a modern industrial estate for the development of a sustainable national industrial competitiveness [36]. In addition, they are also expected to contribute to aspects of sustainability, namely economic, social, and environmental development [36]. The future development of Indonesian industrial estates is also directed at providing services rather than merely selling products [36].

This article will use the PSS concept to help Indonesian industrial estate firms that play important roles in national economic development to gain enough profit and can be sustained in each stage of the life cycle. According to the previous research, actually Indonesian industrial estate firms are starting to do business activities that match with PSS concept [28]. This finding indicates that Indonesian industrial estate firms are suitable for PSS implementation. Therefore, this article will provide a conceptual framework to manage PSS in each stage of the industrial estate firms' life cycle to sustain the industrial estate firms. This article will also provide direction for relevant further research. The conceptual framework also attempts to help the Indonesian government to set a plan, policy, and regulation that will be suitable for the industrial estate firms' development.

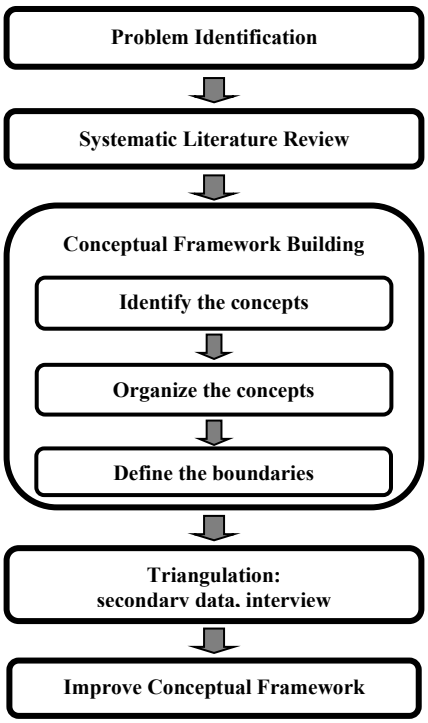


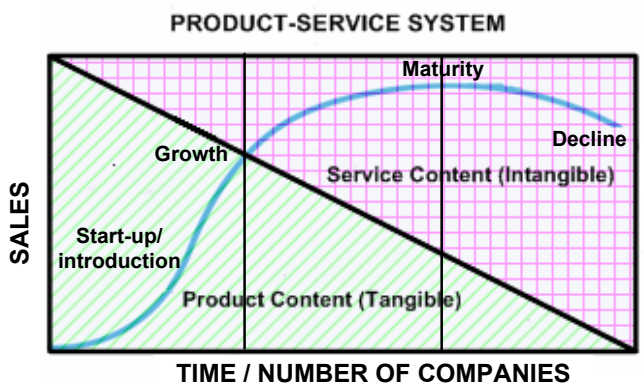
Figure 3. Conceptual Framework Building

To create the conceptual framework, deductive and inductive approach used as shown in Figure 3. A deductive approach will be used to build the framework based on literature reviews while inductive approach will be used to test and improve the framework.

The research begins by identifying problems by examining the condition of industrial estates in Indonesia. Next, authors conduct a systematic literature review on the industrial estate, firm life cycle and PSS to find suitable concepts to build the conceptual framework. The conceptual framework then was built deductively. To build the conceptual framework, there are several methodologies known. This article uses a methodology that suggests three steps of conceptual framework building: (1) identify the concepts, (2) organize the concepts, and (3) define the boundaries [40]. After that, some cases of industrial estate firms will be used to test the conceptual framework. The data will be obtained from secondary data from websites and results. The framework testing will use an inductive approach. The conceptual framework would be improved if needed based on the test result.

To test the conceptual framework, 5 industrial estate firms would be obtained. Four of them are 2015 industrial estate awards winners, namely First Industrial Estate, Second Industrial Estate, Third Industrial Estate, and Fourth Industrial Estate. Industrial estate awards event was conducted in 2015 by Indonesian Industrial Ministry especially to monitor industrial estate development. These industrial estate firms were chosen because they are well-established, good performers and already proven their performance and sustainability for about 30 years. One of the

industrial estate firms taken is the latecomer or new industrial estate firms, namely Fifth Industrial Estate. This industrial estate firm is taken to compare the condition with the four well-established industrial estate firms. Interviews were conducted with experts whose are works in industrial estate firms and consultant of industrial estate firms. The names of the industrial estate firms are disguised in this article, to maintain the confidentiality of the firms.



PRODUCT-ORIENTED PSS	USE-ORIENTED PSS	RESULT-ORIENTED PSS
• Product related ○ Sell products, such as land, buildings, equipments, vehicle ○ Provide maintenance, repair, and supplies services • Advice and consultancy ○ Advice about maintenance and product usage ○ Consultancy about product usage or efficiency	• Product lease ○ Lease products with unlimited access, such as land, buildings, machinery ○ Provide maintenance, repair, and supplies services • Product renting/ sharing ○ Rent and share products with limited access, such as equipments, vehicles ○ Provide maintenance, repair, service • Product pooling ○ Rent products to be used together by customers simultaneously, such as packing equipments, delivery equipments ○ Provide maintenance, repair and service	• Activity management ○ Provide various services that will support customers with outsource system with regular periodical payment, such as security services, cleaning services • Pay per service unit ○ Provide various output of products and services with payment per output taken by companies, such as copy machine, washing machine • Functional result ○ Provide comfort office, company environment, fast delivery, etc

Figure 4. Conceptual Framework of PSS Management

4. Conceptual Framework

As discussed earlier, the problem faced by Indonesian industrial estate firms is how to sustain at every stage of the firm life cycle. Based on this problem, systematic literature reviews were conducted to build the conceptual framework. The first step in building a conceptual framework according to the chosen methodology is to identify the concepts. From the literature review, observation, and interview, it is found that industrial estate companies experience the firm life cycle pattern [41]. At the initial stage of the life cycle, industrial estate firms gain enough profit from selling the product in the form of land and building. But after that, the profit will be decreased because the land and building would be sold out.

Some of Indonesia's long-standing industrial areas maintain their sustainability by expanding and/or creating new business models based on services. The literature review is also carried out to find a suitable concept to help industrial estate companies to be sustainable at each stage of the firm life cycle. The literature review found that the article that links between PSS with the firm life cycle is still rare. In addition, articles on the application of PSS to industrial estate companies have not been found. Therefore, this article will try to apply the concept of PSS related to the life cycle of the company in the Indonesian industrial estate firms. The concept of PSS is found to be appropriate to address this problem, as it offers a new business model that combines products with services. Then it can be concluded that this article will use the concepts of industrial estate firms, firm life cycle, and PSS.

The second step is to organize the concepts. As discussed previously, Indonesian industrial estate firms will generally follow the firm life cycle [41] although the details of the form and the pattern period may be different from the common pattern. Then to be sustained at each stage of the life cycle, the firms need to apply particularly suitable strategies [29]. Considering the characteristics of industrial estate firms that have limited tangible resources, the PSS concept that seeks to shift the use of products to services will be appropriate. The concept of PSS consists of 3 types, each of which consists of different business activities to shift the consumption of products into services [35]. Each of these types of PSS can be implemented as a strategy at each stage of the product life cycle in an attempt to be sustained. For example, for the introduction and growth stages, product-oriented PSS would be suitable, because mostly in these stages industrial estate firms sell products. But this is not the case for other stages. The relationship of the concepts built the conceptual framework for industrial estate firm as described in Figure 4.

This figure shows that in course of their life, industrial estate firm, as also most firms, will face stages of firm life cycle, namely start-up/introduction, growth, maturity, and decline [8]. Due to the limitations of products to be sold, then in the early stages of the firm life cycle, industrial estate firms can sell more products than services. However, in the later stages, when the product is sold out, in order to sustain they need to provide more services. This condition is matching with the characteristics of the PSS types' distribution, namely product-oriented PSS, use-oriented PSS, and result-oriented PSS. Industrial estate firms can appropriately apply product-oriented PSS in the early stages of the product life cycle, where they still have products to sell or product sales are still increasing. After that, use-oriented PSS will be appropriately applied, when the product starts to run out or product sales begin to decline. At last, result-oriented PSS will be appropriate to be applied when the product is fewer or even exhausted and product sales are declining significantly, so new income from the new business model is needed. With the new business, the industrial estate will be survived even in decline stage.

For each type of PSS, there are business activities that can be implemented by the firms. Using the product-oriented type, most firms sell product and added services [35]. Most Indonesian

industrial estate firms' core business is selling products in form of land and buildings. Some industrial estate firms also sell supporting facilities and equipment, but usually, this is not the core business. Service provided by the firms that accompanying product, mainly with purpose to prolong its useful life, in form of:

(1) Product-related services. The business activity where firms sell products to the customer but also add services to prolong the product's useful life, efficient use and take back after usage [35]. Industrial estate firms provide maintenance, repair and supplies services for companies' land, buildings, and facilities, such as the industrial vehicle; machinery; waste; water treatment, etc.

(2) Advice and consultancy. The business activity where firms give advice and consultancy regarding companies operations, management or setup to help prolong the product's useful life and efficient use [35]. In these business activities, the industrial estate firms can provide assistance in form of providing training and consultancy for several activities, such as inventory control, company setting, organizational management, waste treatment, maintenance, etc.

Using use-oriented type, firms sell the use or availability of product [10] instead of sell product [35]. Indonesian industrial estate firms' do business activities of this type in form of :

(1) Product lease. The business activity where firms lease products instead of selling them [35]. Products ownership would still be at producer instead of customer [35]. The producer also has responsibilities of maintenance and taking back products at the end of its useful life [35]. With this business activity, customers get unlimited access to the products leased [35]. Industrial estate firms can lease land; building; industrial vehicles; warehouse; and other equipment to customer/user. Customer/user can use and access the products unlimitedly during the period of leasing. Hence, customer/user is defined as industrial companies operated in the industrial estates firms. Companies, as the customer/user, pay regular fee periodically, while maintenance and repairment ensured by industrial estate firms. This action will ensure cost-efficiency in the whole industrial estate firms due to economic scale and utilization.

(2) Product renting or sharing. The business activity where customer/user rent products and the products can be used by more than one customer [35]. Here, customer/user does not have unlimited access [35]. Industrial estate firms can provide facilities, such as a warehouse, office building, and equipment, such as the vehicle, and machinery to be rent to companies. Customers can share in using these products with other. Industrial estate firms can also rent it to more than one customer with the particular arrangement. This scheme will increase the utilization and productivity of products, decrease total products produce, as well as benefit the customer at an efficiency, practical and lower cost.

(3) Product pooling. The business activity where products can be used together and be operated for more than one customers simultaneously [35]. Industrial estate firms can provide some industrial vehicles, equipment, and facilities that can be used by customers simultaneously. For instance, to deliver production results, more than one customers can use the same vehicle. Another example is water treatment, which more than one customers can use the same equipment. This scheme would be applicable for industrial vehicle and equipment. However, for land, buildings, and specific machinery this scheme will only be applicable for companies that only need the products occasionally or operated in a low capacity.

Using the result-oriented type, the firm sells result or capability instead of products, that can be customized to what customers/users needs [10]. Customers/users will then pay only for the result used [10]. Industrial estate firm can do business activities in form of:

(1) Activity management/outsourcing. The business activity where firms provide outsourcing for customers [35]. Industrial estate firms can provide various services that will support customers to operate more efficiently. Some services that can be provided are cleaning services; garden maintenance; security services, etc. Customers may use those services with the particular amount of payment included in the contract. With this business activities, customers can focus on their own core business, and do not need to waste much energy and effort to manage the supporting activities and the workforce who carry it out.

(2) Pay per service unit. The business activity where firms provide services for customers that are charged per services used by the customer [35]. Industrial estate firms can provide document center; food service; laundry service; load or unload service; and transportation service. Customers can use the services by paying the result according to the number of output needed and used. With this business activity, customers will benefit from efficient cost spent on facilities that actually not continuously or intensely used.

(3) Functional result. The business activity where firms sell functional results instead of products [35]. Industrial estate firms can provide results as services for customers, such as a clean working space; comfortable office; healthy worker; skillful worker, etc. Customers will only pay for the results and get the benefit, without the need to manage or have the skill to do the process by themselves.

The third step of conceptual framework building is to define boundaries. This conceptual framework will be build for Indonesian industrial estate firms because the data were taken from Indonesian industrial estate firms. The basic concept would be compatible for industrial estate firms in other countries but the data must be obtained to adjust whether there are differences in the regulations, conditions, roles, and characteristics of industrial estate firms.

The next step is framework testing to confirm the deductive conceptual framework. Framework testing will be conducted using official websites, papers, and interviews with experts as summarized in Figure 4.

5. Result and Discussion

Indonesian government issuing regulation that industrial estate firms have obligation to provide at least: (1) standard water treatment installation; (2) sewage water treatment installation; (3) drainage; (4) street-lighting; (5) and road for the companies operated in the estate[15]. Then we can conclude that these are the minimum requirement of industrial estate firms. In Indonesia,

according to several sources, it is known that most industrial estate firms depend so much on products selling since it gives large enough profit compares to providing services.

To investigate the implementation of types of PSS and their business activities in Indonesian industrial estate firms, data were obtained from the official websites, the papers, and interviews with experts that closely related to industrial estate firms. As mentioned before, in this article we study 5 industrial estate firms to test the framework. The first four industrial estate firms are 2015 industrial estates award winners whose operated about 30 years, and the fifth is a new industrial estate firm, operated below 5 years. The first three industrial estate firms are located in Java island and the last two are outside Java island.

From our study, we found that all industrial estate firms studied already fulfilled, even exceed, all standards determined by the government. They already provide water supply, water treatment and waste treatment services, security, lighting, road and network, and other infrastructure. The industrial estate firms also found already doing business activities in types of PSS, as summarized in table 1.

From the data, we found that all industrial estate firms studied are focus on selling land and/or buildings as their main business in the early firm life cycle period. Beside selling land and/or buildings, all industrial estate firms also provide maintenance and security services to help customers in operating their businesses. Additionally, the fourth industrial estate provides consultancy and advice services to assist companies (as customers) that will invest in the industrial estate, and assist license application. The fifth Industrial Estate also provides 3 hours investment licensing system to ease companies joining the estate. This fact has proven that Indonesian industrial estate firms in the early stage of life cycle already implement business activities that comply with product-oriented PSS.

Although all industrial estate firms implement product-related business activities, the fourth industrial estate, and the fifth industrial estate also provide implementation advice and consultancy. Among the industrial estate firms, it is found that the kind of services and products provided are almost the same. All of them get the most profit that comes from selling products and provide services in form of maintenance, repair, and security. The fourth industrial estate and the fifth industrial estate have uniqueness in form of license application assistance that maybe become the competitive advantage.

As mentioned in table 1, all industrial estate firms provide warehouses for lease, while the first industrial estate and the second industrial estate also lease factories to customers and the fifth industrial estate rent tank farms. This fact shows that all industrial estate firms already implement product lease business activity with almost the same kind of services. The third industrial estate found rent vehicle, while the fifth industrial estate plan to built meeting incentives, convention, and exhibition (MICE) facility. In many countries, MICE recently believed can attract visitors and bring economic benefit, raising good image, and have a positive impact on the development of the local region as well as the country.

Due to implementing and planning rent business activities, the third and the fifth industrial estate already implement the product rent/sharing business activity. All industrial estate firms provide logistic service with pool system. Additionally, the second and the fifth industrial estate provide a dry port facility that can ease and fast the distribution process. This condition shows that the first to the fourth industrial estate already implement use-oriented PSS through the business activities. Whereas the fifth industrial estate as the considerably new industrial estate has also implement use-oriented PSS, although some activities are still in the planning phase. In use-oriented PSS, it can be seen that there slightly different services provided by industrial estate firms. The differences mainly because of difference in resource and capability owned by the industrial estate firms.

The third industrial estate provides custom office in the estate, while the fourth industrial estate provides manpower management and recruitment service and provide immigration clearance to help companies in the estate. The second industrial estate provides security, office boy and cleaning service outsources, managed by the industrial estate firm. Paying periodical fee, companies in the estate can use these outsourcing services. These services show that the industrial estates already do management business activity, helping customers to do their non-core business with outsourcing system. The first to the fourth industrial estate already do pay per service unit business activities, while the fifth industrial estate still finds the one pay per service business activity. All industrial estates provide hotel. The first to the fourth industrial estate provide health service. The unique pay per services business activity provided by the second industrial estate by renting helicopter; the third industrial estate by providing vehicle service; and the fourth industrial estate by providing dormitory and executive village. Furthermore, the first to the fourth industrial estate already do functional result business activity. The first to the fourth industrial estate provide commercial facilities or shop houses and sport and entertainment facilities, while the first and the second industrial estate build a city around the industrial estate, the second and the third industrial estate build educational facilities, and the second industrial estate built senior living facility. The fifth industrial estate plan to build community academy, but the training already conducted for Sei Mangkei's human resource. From these business activity implementations, it can be concluded that the first to the fourth industrial estate firms already implement result-oriented PSS, while the fifth industrial estate firm still starting to plan to implement result-oriented PSS, and already start to provide training. Based on the analysis, in result-oriented PSS, the difference of services provided even more. These difference can become the competitive advantage for each industrial estate firm.

All types of PSS already found in the first to the fourth industrial estate practice that means services already taken as an important point in their business. However, they were still emphasizing the limited land and building sales, especially during the early stages of their life cycle, since it is providing a great profit. That is why most industrial estate firms attempt to expand the land or find another location once the products sold out. This is due to the fact that selling products still becoming the greatest income generator for industrial estate firms. Nevertheless expanding the industrial estate not always a simple or possible option, because of the high investment and spatial regulation. On the other hand, they still have

Table 1. PSS Implementation in Industrial Estate Firms

Types of PSS	Business activities	First Industrial Estate	Second Industrial Estate	Third Industrial Estate	Fourth Industrial Estate	Fifth Industrial Estate
Product-oriented PSS	Product related	• Sell land and building • Provide maintenance • Provide security	• Sell land and building • Provide maintenance • Provide security	• Sell land and building • Provide maintenance • Provide security	• Sell land and building • Provide maintenance • Provide security	• Sell land and building • Provide maintenance • Provide security
	Advice and consultancy				• Advice and assist companies on incorporation • Advice and assist license application	• Provide assistance with 3 hours of investment licensing service
Use-oriented PSS	Product lease	• Lease warehouses • Lease factories	• Lease warehouses • Lease factories	• Lease warehouses	• Lease warehouses	• Lease warehouses • Lease tank farms
	Product renting/ sharing			• Rent vehicle		• Rent meeting incentives, convention, and exhibition facility (in the plan)
	Product pooling	• Provide Logistic service	• Provide logistic service • Provide dry port facility	• Provide logistic service	• Provide logistic service	• Provide logistic service • Provide dry port facility
Result-oriented PSS	Activity management		• Provide security, cleaning service, office boy outsource	• Provide custom office	• Provide manpower management and recruitment service • Provide immigration clearance	
	Pay per service unit	• Provide hotel • Provide health service	• Provide hotel • Provide health service • Provide helicopter rent	• Provide hotel • Provide health service • Provide vehicle service	• Provide hotel • Provide health service Provide dormitory • Provide executive village	• Provide hotel
	Functional result	• Provide commercial facilities • Provide sport and entertainment facilities • Built city	• Provide commercial facilities • Provide sport and entertainment facilities • University • Built city • Provide senior living facility	• Provide shophouses • Provide sport and entertainment facilities • Vocational high school	• Provide commercial facilities • Provide sport and entertainment facilities	• Provide training as an impetus of community academy

Source: summarized from official websites [42]–[45], a conference material [46]and interviews with experts

obligations to their customers as in the regulation that they still must manage the industrial estate and provide service to the companies inside. Therefore, they manage to provide various services to be sustained. The more they mature they will find more effort to innovate to provide more various services.

The first to the fourth industrial estate was built in the 1990s, range from 1989 until 1993. It means that they have operated more than 30 years and already in the middle or later stages of the life cycle. This is one of the reasons why their business activities tend to innovate in providing services. By providing more services,

industrial estate firms have more opportunity to do innovation. With innovation, firms have more opportunity to gain new profit to be sustained. In contrast to the four industrial estates that has long been operating, the fifth industrial estate has only been operating for about 3 years and still at the early stage of firm life cycle. So the fifth industrial estate still provides a little kind of service or still planning the services. This conforms with the framework that the more mature the industrial estate the more they need to provide service. It can be said that use-oriented PSS and result-oriented PSS would be needed when an industrial estate firm going to runs out their land and buildings to keep on sustain.

Among the industrial estate firms, while implementing product-oriented PSS, they mainly sell and provide similar products and services with similar business activities. Nevertheless, by using use-oriented PSS and moreover, result-oriented PSS, which is the composition of providing services more than products, among industrial estate firms there are differences in business activities and services provided. For example, there are industrial estate firms that rent factories, build a city, provide sport and entertainment facilities while the other create services such as healthcare, immigration clearance, etc. Then, it can be concluded that by providing more services, industrial estate firms can do more various business activities.

From the conceptual framework and discussion, there are propositions that can be built regarding PSS implementation in industrial estate firms toward sustainability, considering the firm life cycle.

Proposition 1: Industrial estate firm would need to provide more services in the later stage of the firm life cycle to be sustained because they have limited tangible products.

Proposition 2: PSS concept supposed to be suitable to be implemented for industrial estate firms, due to the emphasizes on services providing rather than tangible products.

Proposition 3: Among types of PSS, product-oriented, use-oriented product-oriented composition of PSS is getting bigger, leading to more possibilities for innovation, as services are given more opportunity for it.

Proposition 4: More service composition leads to the industrial estate firms that have more flexibility to create business activities. That is why in the application of results-oriented PSS, there are many variations in the types of services provided between industrial estate firms. This variation can then become a uniqueness that will form a competitive advantage for industrial estate firms.

Proposition 5: In relation to Proposition 2, industrial estate firms with unique business activities and competitive advantage based on services will have the greater likelihood of profit and sustainability.

Proposition 6: All three types of PSS are appropriate to be applied to each product life cycle in order to maintain the sustainability of the industrial estate firm at each stage of the product life cycle. Product-oriented appropriate PSS is applied in the early stages of the company's life cycle when the firm is new and sells many products in the form of land and buildings. The appropriate use-oriented PSS is applied at the mid-life stage of a company when the product begins so little that the service has begun to be emphasized. Result-oriented PSS is appropriate to apply to the end of the firm life cycle when the product is getting smaller or exhausted, so the service needs to be the main business activity.

6. Conclusion

As already been explained, this article is extending the previous work in 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) [1]. In this article, we added the concept of firm life cycle and data from interviews with experts. The novelty of this article is by added firm life cycle concept in PSS management. This concept is added because of the knowledge regarding industrial estate firms will follow a firm life cycle pattern [41]. As discussed earlier, that would create differences with the condition [28] and strategy needed [29]. Therefore, different PSS management would be promptly needed.

From the above discussion, it can be concluded that the concept of PSS can be applied to help Indonesian industrial estate firms to be sustainable across their life cycle. This is strengthened by data and interview results that many Indonesian industrial estate firms have started to think and provide services combined with products sales. Among firms, there are differences of services provided and business activities done. These findings are in line with the reference that state industrial estate firms need to shift to provide service for further development to be sustained [36].

Facing the firm life cycle, industrial estate firms can use the three types of PSS with the business activities. Industrial estate firms can do business activities that are suitable for their conditions regarding with firm life cycle. The type and business activities can be changed when the condition of the firm change along the firm life cycle. The suitable PSS management would lead to firms' sustainability in each life cycle. The conceptual framework suggests that the types of PSS implementations can assist industrial estate firms at each stage of the firm life cycle through creative innovation of the services provided to enable the industrial area to play its role well.

From the data, it is also found that the more the service composition is provided, the industrial estate firm will have more flexibility to innovate creatively. The more innovation would lead to more opportunity creation and the ability to choose the right business and service type. With this benefit, the industrial estate firms would get a competitive advantage, not to mention sustainability.

The implication of the findings is that industrial estate firms need to manage their PSS as an attempt to be sustained. They better plan the PSS in the beginning for they can get more stable and continuous profit. The planning for PSS from the beginning will also ease industrial estate firms to allocate the service facilities in the industrial estate and maintain good integration in the estate.

This article will help industrial estate firms to see opportunities for business activities that meet the conditions that can be applied at each stage of the firm life cycle based on the PSS concept. This concept also provides suggestions for industrial estate firms in determining strategies at each stage of the firm life cycle to be sustainable. For the government, this article will support the

government's efforts to develop appropriate policies in the development of the industry through the role of the industrial estate, which in turn improves the national economy and welfare. In the academic field, this article presents the incorporation of industrial estate concepts, corporate life cycles and PSS that are currently unexamined. This conceptual framework can generally be used for industrial estate firms because the basic theory used is general. However, since the empirical data in this article are only taken from industrial estate firms in Indonesia, there may be differences in government regulations, and conditions of competition, business, economy, industry conditions, etc for other countries that may require adjustment.

The implementations of PSS in this article are identified from the websites of industrial estates award winners and the interviews with industrial estate firms experts. For a better study, data from the websites need to be triangulated to get better descriptions. In addition, for further research, it is necessary to study more industrial estates to strengthen the results obtained. The study with non-award winners industrial estate firms also needs to be conducted to study to further test the concept. Moreover, the new and middle age industrial estate firms also need to be studied, to confirm the concept. Then our next research is to test the conceptual framework that has been built with interviews with several industrial estate firms with different conditions.

Further research about the comparison between private and public industrial areas as well as the comparison between industrial areas located in Java and those located outside Java should be employed. Research can also be extended by combining other concepts of operational strategies. Another research can also be conducted to study the firm life cycle pattern. Further study would also be needed to study the industrial estate firms with various time operated to better see the pattern of implementation of each type of PSS in industrial firms.

Conflict of Interest

The authors declare no conflict of interest

Acknowledgment

First author thanks the Education Fund Manager Institution (Lembaga Pengelola Dana Pendidikan/LPDP) in The Ministry of Finance, and The Ministry of Research, Technology and Higher Education, Republic of Indonesia for funding this research. Authors also thank Ms. Prameshwara Anggahegari for proofreading this manuscript.

References

[1] C. Wirawan, G. Yudoko, and Y. D. Lestari, 'Product-service system for Indonesian industrial estate firms : A conceptual framework', in *Proceedings of 2017 IEEE IEEM*, Singapore, 2017, 1812–1816.https://doi.org/10.1109/IEEM.2017.8290204

[2] P. Scott, 'Industrial Estates and British Industrial Development, 1897-1939', *Bus. Hist.*, 43(2), 73–98, 2001.http://dx.doi.org/10.1080/713999223

[3] A. R. R. Ramos and F. P. Fonseca, 'A methodology to identify a network of industrial parks in the Ave Valley, Portugal', *Eur. Plan. Stud.*, 20(10), 1844–1862, 2016.http://dx.doi.org/10.1080/09654313.2016.1202201

[4] World Bank Group, *Getting to Green - A Sourcebook of Pollution Management Policy Tools for Growth and Competitiveness*. www.worldbank.org, 2012.

[5] F. Fonseca, R. A. R. Ramos, and A. N. R. da Silva, 'An agent-based model to assess the attractiveness of industrial estates', *Jasss*, 18(4), 1–11, 2015.DOI: 10.18564/jasss.2893

[6] United Nations Industrial Development Organization, 'Economic Zones in the ASEAN: Industrial parks, special economic zones, eco-industrial parks, innovation districts as strategies for industrial competitiveness', August. 2015. DOI:10.1016/j.indmarman.2016.03.001

[7] B. R. Scott, *Stages of corporate development*. Boston: Harvard Business School Press, 1971.

[8] H. Mintzberg, 'Power and Organization Life Cycle', *Acad. Manag. Rev.*, 9(2), 207–224, 1984.

[9] O. K. Mont, 'Clarifying the concept of product-service system', *J. Clean. Prod.*, 10(3), 237–245, 2002. https://doi.org/10.1016/S0959-6526(01)00039-7

[10] T. S. Baines, H. Lightfoot, E. Steve, A. Neely, R. Greenough, J. Peppard, R. Roy, E. Shehab, A. Braganza, A. Tiwari, J. Alcock, J. Angus, M. Bastl, A. Cousens, P. Irving, M. Johnson, J. Kingston, H. Lockett, V. Martinez, P. Michele, D. Tranfield, I. Walton, and H. Wilson, 'State-of-the-art in product service-systems', in *Proceedings IMechE, 221 Part B: J. Engineering Manufacture*, 221, 1543-1552, 2007.DOI: 10.1243/09544054JEM858

[11] A. P. B. Barquet, M. G. de Oliveira, C. R. Amigo, V. P. Cunha, and H. Rozenfeld, 'Employing the business model concept to support the adoption of product-service systems (PSS)', *Ind. Mark. Manag.*, 42, 693–704, 2013.http://dx.doi.org/10.1016/j.indmarman.2013.05.003

[12] CDMI Consulting Group, 'Studi potensi bisnis dan pelaku utama kawasan industri di Indonesia 2016-2020', Jakarta, 2015.

[13] Republic of Indonesia Industrial Ministry, *Rencana Induk Pembangunan Industri Nasional 2015-2035 (National Industrial Master Plan 2015-2035)*. 2015.

[14] Republic of Indonesia., 'Undang-Undang Republik Indonesia Nomor 3 Tahun 2014 Tentang Perindustrian (Republic of Indonesia Law no. 3 the year of 2014)'.

[15] Republic of Indonesia. Industrial Ministry., 'Peraturan Pemerintah Republik Indonesia Nomor 142 Tahun 2015 Tentang Kawasan Industri (Republic of Indonesia Government Regulation No. 142 the year of 2015 about industrial estate)'.

[16] G. Yudoko, 'Strategy map for industrial estate firms: model building in the Indonesia Regulatory Context', *Adv. Sci. Lett.*, 22(12), 4316–4320, 2016.https://doi.org/10.1166/asl.2016.8136

[17] G. Yudoko, 'The Indonesia Industrial Estate Firm Award's: Theory Building and Testing', *Soc. Sci.*, 11(6), 7341–2345, 2016.

[18] J. Elkington, 'Enter the Triple Bottom Line', in *The Triple Bottom Line: Does It All Add up?* A. Henriques and J. Richardson, Eds. London: Earthscan, 2004, 1–16.

[19] E. Dewberry, M. Cook, A. Angus, A. Gottberg, and P. Longhurst, 'Critical Reflections on Designing Product Service Systems', *Des. Journal*, 6(4), 408–430, 2013.DOI: 10.1016/j.compind.2012.02.006

[20] S. Wiesner, M. Freitag, I. Westphal, K.-D. Thoben, and X. Boucher, 'Interactions between service and product lifecycle management', in *Procedia CIRP*, 2015, 221, 36–41.DOI: 10.1016/j.compind.2012.02.006

[21] S. Cavalieri and G. Pezzotta, 'Product-service systems engineering: State of the art and research challenges', *Comput. Ind.*, 63, 278–288, 2012.DOI: 10.1016/j.compind.2012.02.006

[22] V. Zanetti, S. Cavalieri, and G. Pezzotta, 'Additive Manufacturing and PSS: a Solution Life-Cycle Perspective', *IFAC-PapersOnLine*, 49(12), 1573–1578, 2016.DOI: 10.1016/j.ifacol.2016.07.804

[23] S. Singhal and A. Kapur, 'Industrial estate planning and management in India - An integrated approach towards industrial ecology', *J. Environ. Manage.*, 66, 19–29, 2016.DOI: 10.1006/jema.2002.0571

[24] V. Veleva, S. Todorova, P. Lowitt, N. Angus, and D. Neely, 'Understanding and addressing business needs and sustainability challenges : lessons from Devens eco-industrial park', *J. Clean. Prod.*, 87, 375–384, 2015.http://dx.doi.org/10.1016/j.jclepro.2014.09.014

[25] C. K. Prahalad and V. Ramaswamy, 'Co-creating unique value with customers', *Strateg. Leadersh.*, 32(3), 4–9, 2004.

[26] I. M. Jawahar and G. L. McLaughlin, 'Toward a descriptive stakeholder theory: An organizational life cycle approach', *Acad. Manag. Rev.*, 26(3), 397–414, 2001.https://doi.org/10.5465/amr.2001.4845803

[27] D. C. Mueller, 'A Life Cycle Theory of the Firm', 20(3), 199–219, 2013.DOI: 10.2307/2098055

[28] C. R. Anderson and C. P. Zeithaml, 'Stage of product life cycle, business strategy, and business performance', *Acad. Manag. J.*, 27, 5–24, 1984.https://doi.org/10.5465/255954

[29] S. H. Hanks, C. J. Watson, E. Jansen, and G. N. Chandler, 'Tightening the life-cycle construct : A taxonomic study of growth stage configurations in organizations', *Entrep. Theory Pract.*, 18(2), 5–30, 1993.https://doi.org/10.1177/104225879401800201

[30] A. Chandler, 'Strategy and structure'. Cambridge. MA, MIT Press, 1962.
[31] R. E. Quinn and K. Cameron, 'Organizational life cycles and shifting criteria of effectiveness: some preliminary evidence', *Manage. Sci.*,29(1), 33–51, 1993.https://doi.org/10.1287/mnsc.29.1.33
[32] M. Kuncoro and S. Wahyuni, 'FDI impacts on industrial agglomeration: the case of Java, Indonesia', *J. Asia Bus. Stud.*, 3, 65–77, 2009.https://dx.doi.org/10.1108/03068290910921226
[33] T. Firman, 'Rural to urban land conversion in Indonesia during boom and bust periods', *Land use policy*, 17(1), 13–20, 2000.DOI: 10.1016/S0264-8377(99)00037-X
[34] M. J. Goedkoop, C. J. van Halen, H. R. te Riele, and P. J. Rommens, 'Product Service systems, Ecological and Economic Basics', PricewaterhouseCoopers N.V. / Pi!MC, Storm C.S., Pre consultants, 1999.
[35] A. Tukker, 'Eight types of product-service system: Eight ways to sustainability? Experiences from suspronet', *Bus. Strateg. Environ.*, 13, 246–260, 2004.DOI: 10.1002/bse.414
[36] D. Mulyadi, *Manajemen Perwilayahan Industri*, First. Jakarta, Indonesia: Leuser Cita Pustaka, 2012.
[37] M. Lindahl, E. Sundin, and T. Sakao, 'Environmental and economic benefits of Integrated Product Service Offerings quantified with real business cases', *J. Clean. Prod.*, 64, 288-296, 2014. DOI: 10.1016/j.jclepro.2013.07.047
[38] S. M. Lee, D. L. Olson, and S. Trimi, 'Co-innovation : Convergenomics, collaboration, and co-creation for organizational values', *Manag. Decis.*, 50(5), 817–831, 2012. DOI: 10.1108/00251741211227528
[39] Y. Geum and Y. Park, 'Designing the sustainable product-service integration: A product-service blueprint approach', *J. Clean. Prod.*, 19, 1601–1614, 2011.DOI: 10.1016/j.jclepro.2011.05.017
[40] R. A. Swanson and T. J. Chermack, *Theory Building in Applied Disciplines*. San Francisco: Berret-Koehler Publishers, Inc., 2013.
[41] F. Belussi and S. R. Sedita, 'Life Cycle vs. Multiple Path Dependency in Industrial Districts', *Eur. Plan. Stud.*, 17(4), 505–528, 2009.DOI: 10.1080/09654310802682065
[42] 'Karawang International Industrial City'. [Online]. Available: www.kiic.com. [Accessed: 09-Jun-2017].
[43] 'Jababeka & Co. Industrial'. [Online]. Available: www.jababeka.com. [Accessed: 09-Jun-2017].
[44] 'MM2100 Industrial Town'. [Online]. Available: www.mm2100.ac.id. [Accessed: 09-Jun-2017].
[45] 'Batamindo Investment Cakrawala'. [Online]. Available: www.batamindoindustrial.com. [Accessed: 09-Jun-2017].
[46] Nurhidayat, 'Kawasan Ekonomi Khusus (KEK) Sei Mangkei Menuju Eco-Industrial Park', *National Seminar on Palm Oil Insight Panel Speech*. Yogyakarta, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 129-141 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Method for Generating, Evaluating and Comparing Various System-level Synthesis Results in Designing Multiprocessor Architectures

Peter Arato*, Gyorgy Racz

Department of Control Engineering and Information Technology, Budapest University of Technology and Economics, 1117, Hungary

ARTICLE INFO

Article history:
Received: 08 May, 2018
Accepted: 02 June, 2018
Online: 15 June, 2018

Keywords:
system level synthesis,
multiprocessor system,
heterogeneous architectures,
decomposition,
high level synthesis,
pipelining

ABSTRACT

Multiprocessing can be considered the most characteristic common property of complex digital systems. Due to the more and more complex tasks to be solved for fulfilling often conflicting requirements (cost, speed, energy and communication efficiency, pipelining, parallelism, the number of component processors, etc.), different types of component processors may be required by forming a so called Heterogeneous Multiprocessing Architecture (HMPA). The component processors of such systems may be not only general purpose CPUs or cores, but also DSPs, GPUs, FPGAs and other custom hardware components as well. Nevertheless, the system-level design process should be capable to handle the different types of component processors the same generic way. The hierarchy of the component processors and the data transfer organization between them are strongly determined by the task to be solved and by the priority order of the requirements to be fulfilled. For each component processor, a subtask must be defined based on the requirements and their desired priority orders. The definition of the subtasks, i.e. the decomposition of the task influences strongly the cost and performance of the whole system. Therefore, systematically comparing and evaluating the effects of different decompositions into subtasks may help the designer to approach optimal decisions in the system-level synthesis phase. For this purpose, the paper presents a novel method called DECHLS based on combining the decomposition and the modified high level synthesis algorithms. The application of the method is illustrated by redesigning and evaluating in some versions of two existing high performance practical embedded multiprocessing systems.

1. Introduction and Related works

This paper is an extension of the work originally presented in IEEE System on Chip Conference 2017 [1]. This extension contains a more detailed explanation of the applied system level synthesis method. An additional practical benchmark is also solved and evaluated.

The system-level synthesis procedure of high-performance multiprocessor systems usually starts by some kind of a task description regarding the requirements and their desired priorities. The task description is usually formalized by a dataflow-like graph or by a high level programming language [2] [3]. To each component processor of the multiprocessor system, a subtask is assigned basically by intuition considering various special requirements (communication cost, speed, pipelining, etc.). The definition of the subtasks, i.e. the decomposition of the task strongly influences the cost and performance of the whole multiprocessing system. Therefore, comparing and evaluating the effects of different task decompositions performed by applying systematic algorithms may help the designer to approach the optimal decisions in the system-level synthesis phase [4]. Our proposed method can also be considered a design space exploration as shown in [3-5]. In [5] the hardware-software partitioning problem is analyzed and various solutions are compared and evaluated. This approach refers formally only to special two-segment decompositions. These can be extended for multi-segment cases by utilizing several high performance heuristic algorithms [6], [7]. However, in case of such extensions for multiprocessing architectures [8], the decomposition algorithm and the calculation to find beneficial communication cost and time

* Corresponding Author: Peter Arato, Email: arato@iit.bme.hu

become rather complex and cumbersome. Other special aspects of multi-segment partitioning problems can be found in [9] at the synthesis of low-cost system-on-chip microcontrollers consisting of many integrated hardware modules. For avoiding to the aforementioned difficulties, the paper presents a novel method called DECHLS based on combining decomposition and modified high level synthesis algorithms.

This paper is organized as follows. Section 2 contains the overview of the proposed novel system-level synthesis method (DECHLS). The decomposition algorithms applied in the experimental DECHLS tool and the applied performance metrics are presented in Section 3. The embedded multiprocessing systems to be redesigned by the experimental DECHLS tool, the redesigned variations obtained by DECHLS and the evaluation of the results are presented in Section 4. The conclusion is summarized in Section 5.

2. Overview of the System-Level Synthesis method DECHLS

Our method (DECHLS [1]) is based on combining decomposition (DEC) with high level synthesis (HLS) algorithms. One of the main steps in HLS tools [10], [11] is the allocation executed after scheduling the elementary operations. The aim of scheduling is to determine a beneficial start time for each elementary operation to attempt allocating them optimally into processing units [12]. Usual optimality criterions in HLS tools are the minimal number of processing units, the lowest cost and the highest speed in pipelining [10], [11], or the lowest power consumption [12]. The elementary operations allocated into the same processing units compose the subtasks. However, such subtasks provided by the HLS tools are strongly determined by the usual scheduling algorithms and so, not enough freedom is left for satisfying special multiprocessing requirements in further synthesis steps. Instead of modifying and extending the usual HLS scheduling algorithms [13-16] for fulfilling these special requirements, our DECHLS method attempts to satisfy these on a higher priority level by executing the decomposition into subtasks already before the HLS phase. A similar approach is presented in [17] called clustering. In [17] the power efficiency is attempted to be optimized by clustering in pipeline systems at a given throughput. The resulting structure can be considered a special decomposition that may decrease communication and hardware costs as well, but these benefits are not analyzed directly. Our preliminary decomposition can be considered a kind of a preallocation providing more freedom to fulfill special multiprocessing requirements. Based on the subtasks (called segments hereinafter) obtained by decomposition, a segment graph (SG) can be constructed. The SG can serve as an elementary operation graph (EOG) of the modified HLS tool. Thus, the scheduling and allocation algorithms in the modified HLS tool refer to the component processors in further steps. In this way, the decomposition executed before the HLS steps can influence the result on a higher priority level. Thus, DECHLS is based on combining the decomposition (DEC) and modified high level synthesis (HLS) algorithms. The main contribution of the DECHLS is that these two steps are integrated into the same design tool enabling the designer to generate, evaluate and compare various alternative results. The list of user-defined requirements, required inputs and outputs of the DECHLS process is outlined in Fig.1. The workflow of the DECHLS method is shown in Fig 2.

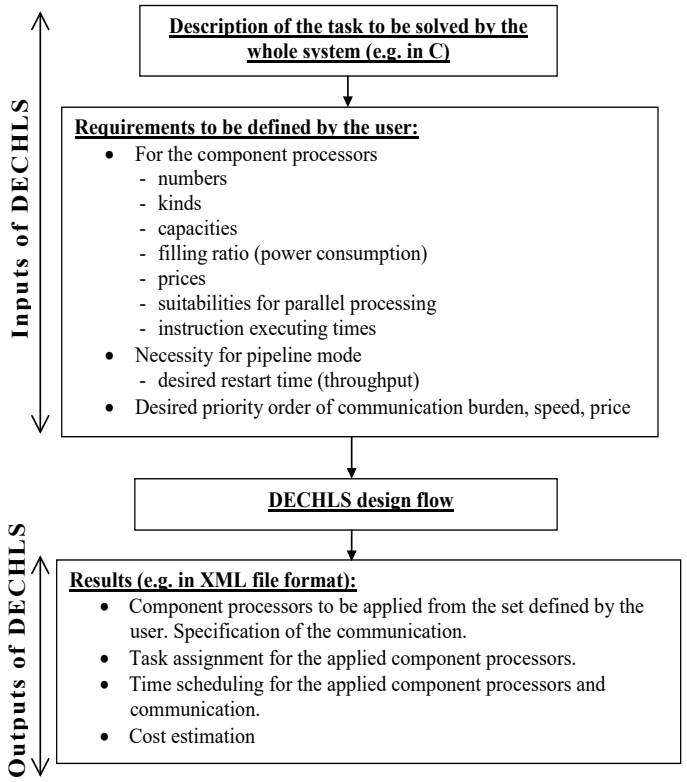


Fig. 1. Inputs and outputs of the DECHLS method

There is also a possibility in DECHLS method to generate the EOG for the HLS tool immediately from the task description without preliminary decomposition. If decomposition is intended to be made, then the desired number of segments can either be prescribed or calculated by the decomposition algorithm. In both cases, the modified HLS tool receives the segment graph (SG) as EOG. If the user decision is that pipeline mode is not required, then the process of DECHLS is finished and the allocated nodes formed from the segment graph (SG) define the tasks for the component processors. If however, the pipeline mode is desired, then the applicable value of the pipeline initialization interval called further on restart time (R) is calculated. Most of the HLS tools [10], [11] can attempt to ensure a user-defined desired restart time (Rd.). If $R \leq Rd$, is not the case, then the HLS tool [10] applied in the experimental DECHLS version reduces R by buffer insertion and/or by applying multiple copies of nodes. If the cost of the resulting multiprocessing system is not acceptable, then allowing a longer latency time L (the longest time of processing an input data in the allocated segment graph) in pipeline mode may reduce the cost by calculating a new schedule [18]. The user can increase L in more iteration steps and the value of dL is also definable by the user. If the cost is acceptable, then the process of DECHLS is finished.

The scheduler applied in the experimental DECHLS version is a modified force-directed one [19]. Results of the usual force-directed schedulers strongly depend on the order of fixing the operations within their mobility domain (i.e. between the earliest and latest possible start time) [20]. In contrary, our modified

scheduler generates a list of all operations according to the length of their mobility and the operations having the shortest mobility are fixed first (i.e. the top of the mobility list). After fixing an operation, the mobility list is recalculated and a next operation is chosen, and so on until each operation has been fixed.

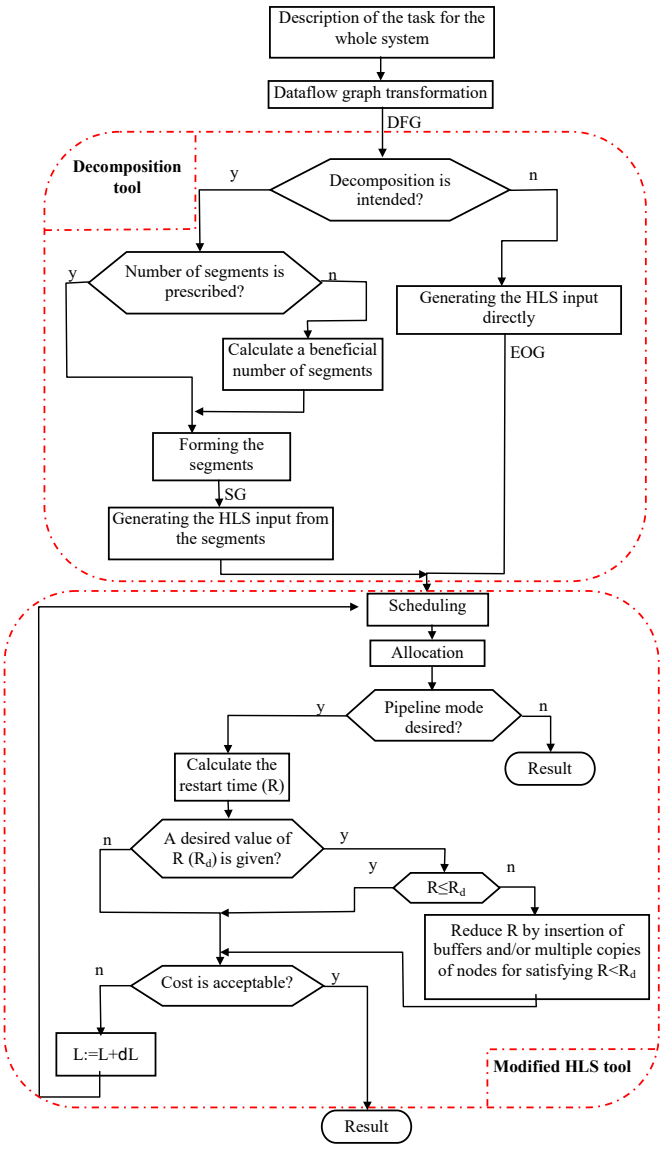


Fig. 2. Design flow of the DECHLS method

Increasing the latency (L) can be simply implemented in any HLS tool. The simplest way is to introduce an extra fictive operation parallel with the original operation graph. The execution time of this extra operation is assumed to be equal to the desired latency value. This extra operation receives all inputs of the original operation graph, and its output enables all outputs of the original graph. This extra operation could be added only before scheduling, and it must be removed before allocation, since it is not an operation to be executed by the resulting system.

3. Decomposition of the Task Description

The decomposition step of the DECHLS method starts from a dataflow-like task description (DFG). The nodes of this graph represent inseparable operations, and the edges symbolize data dependencies between the operations. The decomposition algorithm should define disjoint segments containing each of them several (at least one) operations. The segments represent subtasks of the task description and form a segment graph (SG) that may be the input of the HLS part in the DECHLS system. The edges in SG represent data communication between segments, edges are weighted by the communication burden. The nodes in SG representing the subtasks are also weighted by the longest execution time of the subtask. The execution time of a segment is the sum of execution times of the elementary operations assigned into this segment. The reason of this is that the processors are assumed to execute serially all elementary operations to be assigned to them. For comparing and evaluating the effect of different task decompositions, the desired number of segments is user-definable. The distribution of workload between segments is arranged beneficially for pipeline implementation (i.e. as far as possible equally) and at the same time, the communication load (edge weights) between the segments should also be kept possibly low. Beside these aims, the decomposition algorithm must not produce segment cycles as shown in Fig. 3 and [21]. Such cycles may cause deadlocks and it is difficult to handle them in pipeline mode. Handling or eliminating the cycles depends on the HLS tool applied in DECHLS system [10] [11]. The decomposition algorithms may produce cycles in the SG, even if the task description (EOG) is acyclic. Therefore, only such decomposition algorithms are allowed in DECHLS system, which guarantee cycle-free SG-s [21]. In our experimental DECHLS tool, this difficulty is avoided by generating the allowable edge cutting places before starting the decomposition algorithm [21].

The decomposition algorithms may yield segments consisting of extremely different number of elementary operations. In the further steps, the allocation algorithm in DECHLS may provide processors with extremely different workload.

Therefore, the user has to decide, whether applying always the just fitting processors or choosing equally suitable identical processors. A very simple cost calculation and comparison (i.e. the number of processors) can be applied in the latter case. Therefore, identical processors are assumed in the further discussions and in the solutions of the practical examples.

A set of decomposition algorithms is available in [21-23]. Two of them seemed to be the most suitable for DECHLS: the spectral clustering (SC) and the multilevel Kernighan–Lin (KL) method. The spectral clustering is applied for the illustrative redesigning examples in two versions: the number of segments is prescribed or it is calculated by the algorithm. These decomposition algorithms are suitable to illustrate the DECHLS method and yield acceptable results. However, other more advanced or flexible decomposition methods like [25] or [26] could also be applied in DECHLS.

For comparing and evaluating the solutions obtained by DECHLS, some performance metrics could be assumed. These metrics are flexibly definable and extendable by the user. The experimental metric criteria applied in Section 4 for comparing the results are as follows. The hardware cost is calculated simply as the number of processors, assuming that each processor has the same performance specification.

- The pipeline restart time versus the hardware cost is plotted for helping the beneficial and economical choices and forecasting the consequences.

- The longer latency in pipeline mode [18] [29] may reduce the hardware cost without significant effects on the restart time.

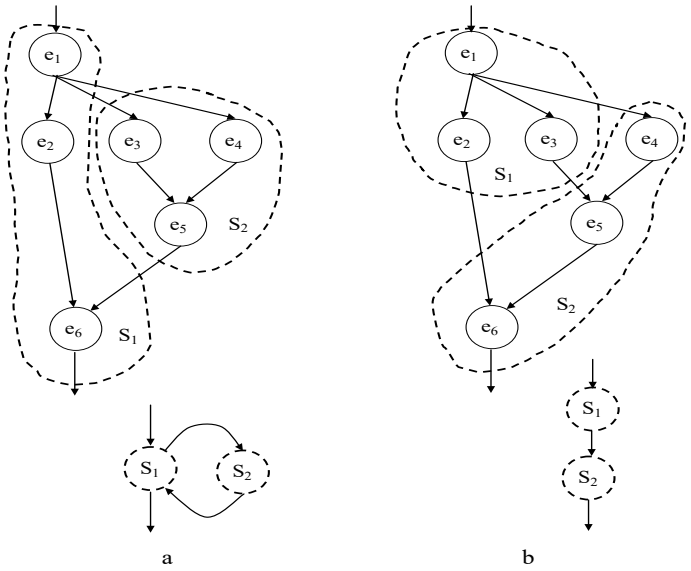


Fig. 3. Allowable (b) and not allowable (a) decomposition [19]

4. Benchmark Solutions and Evaluations

4.1. A sound source localization system

A sound source localization system is presented in [27] as a relatively simple multiprocessing system. The segment graph (SG) constructed probably by intuition is shown in Fig. 4.

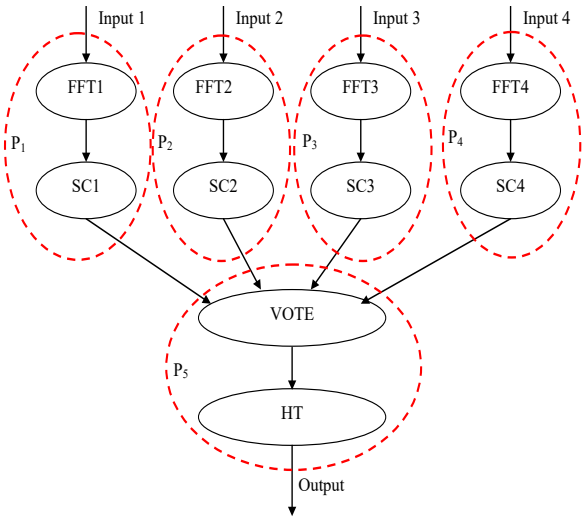


Fig. 4. Intuitively decomposed and allocated segment graph (SG) (sound source localisation) [27]

The task to be solved is decomposed into 10 subtasks (each of them is a separate segment) FFT1, … FFT4, SC1, … SC4, VOTE, HT as shown in Fig. 4. The segment functions are as follows. FFTs execute a 512 point real fast Fourier-transformation. SCs perform noise power estimation for each channel of input data. VOTE detects whether a detectable sound was present on at least two channels in the current sample, and the noise is sufficiently low at the same time. The longest segment is HT that performs the hypothesis testing method [27] for determining the source coordinates of the sound.

The implementation in [27] is based on five processors (P1, ... P5) as shown in Fig. 4. One of the consequences of this intuitive solution is that P5 has the computationally most expensive subtask requiring approximately 5 times faster processor than the other ones. Therefore, the speed of the whole system strongly depends on the implementation of P5.

In order to evaluate new decompositions, the given segment graph has to be expanded into an elementary operation graph (EOG), the nodes of which are considered atomic inseparable elementary operations. A possible method may start with a high level language description of the segments in SG. Each segment may correspond to a function in the high level language description. This function may call other functions and so on until getting functions considered inseparable.

For the sake of simplicity, software loops in the task description can also be considered single inseparable nodes, because the complex loop handling would be an additional difficulty (e.g. in pipeline systems [28]). This simplification does not affect the essence of the further steps.

The transformation process from SG to EOG is illustrated by a simple example in Fig. 5.

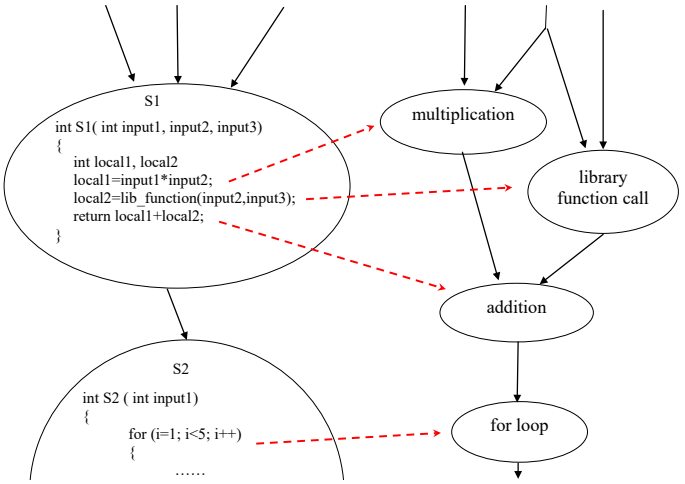


Fig. 5. Expanding a SG into an EOG
(the small program fragments in segments are only to illustrate the process)

From the SG in Fig. 4, this transformation yields an EOG with over 160 nodes.

For further handling of the EOG, execution times should be assigned to each elementary operation. For these assignments, the specifications of processors to be applied should be known. In the original implementation [27], two different processor types have been used: the slower MSP430 and the faster ARM7. The faster ARM7 processor has been chosen probably because of the large memory demand of HT. The slower and cheaper processors, (16 bit MSP430 microcontrollers) have been found suitable for the remaining tasks.

Further on, the execution times will be expressed by the number of time steps as used in our modified HLS tool [10]. In this way, the calculation becomes independent from the clock speed. All time data, including data of the existing systems will also be expressed in time steps for comparison.

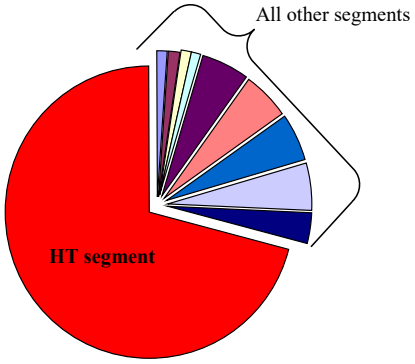


Fig. 6. The relative execution time proportions of the segments in Fig. 4.

Fig. 6 shows the relative execution time values of the segments determined according to the intuitive decomposition shown in Fig. 4. Note, that this solution yielded great differences in the execution times of the processors that may cause unnecessary constraints in case of pipelining.

The comparative solutions obtained by applying DECHLS approach to a uniform workload distribution between the processors.

4.2. Evaluation of the existing decomposition for the sound source localization

As it has been mentioned earlier, only identical processors are assumed in both redesign examples. In order to achieve the best results with identical processors, all the segments require approximately identical execution times. To illustrate the problem, our modified HLS tool provides the plot in Fig. 7. calculated from the intuitive SG in Fig. 4. The execution times are calculated assuming that only MSP430 microprocessors are applied in the implementation. Such an implementation will be considered as comparison in the further solutions. This assumption helps to illustrate the effect of different preliminary decompositions without complicating the comparison of different results.

Fig. 7 does not show a decrease after time step 2300, thus at least 11 identical processors would be required for this solution. The reason of this is that the processor executing the HT needs to be replicated at least in two copies in order to meet the timing constraints.

4.1. Results provided by DECHLS without decomposition for the sound source localization

In this mode of DECHLS, the HLS tool (Fig. 2) receives directly the EOG without preliminary decomposition, i.e. the decomposition is left completely for the allocation algorithm of the HLS tool. It can be observed in Fig. 8 that after the restart time R=3000 no further cost decrease occurs in this mode of DECHLS. The cost minimum is here 14 processors instead of 11 obtained by the intuitive decomposition (Fig. 7). In this calculation, the HLS part of DECHLS has been finished as if the cost would be acceptable (Fig. 2). Therefore, the latency has not been increased. In this case, the latency determined by the EOG is 4217 time steps. To the solution in Fig. 8 belongs a latency value of 5025 time steps.

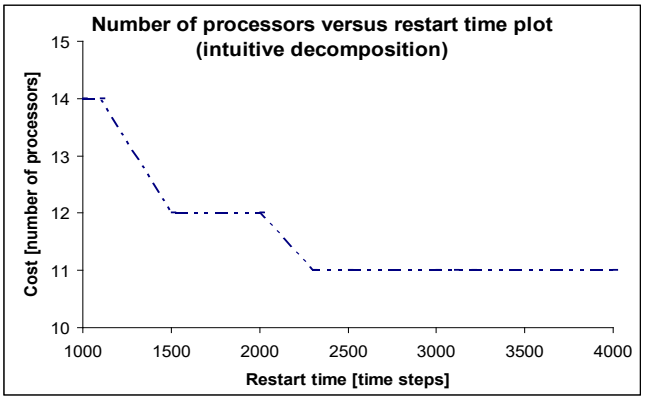


Fig. 7. Number of processors versus restart time by intuitive decomposition (sound source localisation)

The cost could be attempted to be reduced both by increasing the latency and by applying suitable decompositions. Some characteristic cases will be shown later.

The runtime of the HLS tool was extremely long in this mode (requiring about 18 hours on a contemporary PC), mostly because of the high number of elementary operations. The longest time has been taken by the force directed scheduling algorithm.

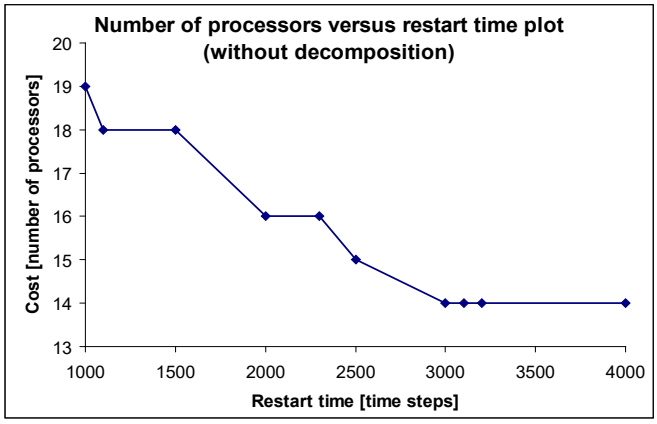


Fig. 8. Number of processors versus restart time without decomposition (sound source localisation)

4.2. Results provided by DECHLS with preliminary decomposition for the sound source localization

In this experiment, DECHLS is applied using both task decompositions [21] based on the multilevel Kerrigan-Lin (KL) [23] algorithm and the spectral clustering (SC) [24]. The results are nearly the same. In both calculations, the desired number of segments has been prescribed as 5 and 7 respectively. Both results are illustrated in Fig. 9 and Fig. 10. In these figures and in the following ones, the relative segment execution times are also shown in small circle diagrams.

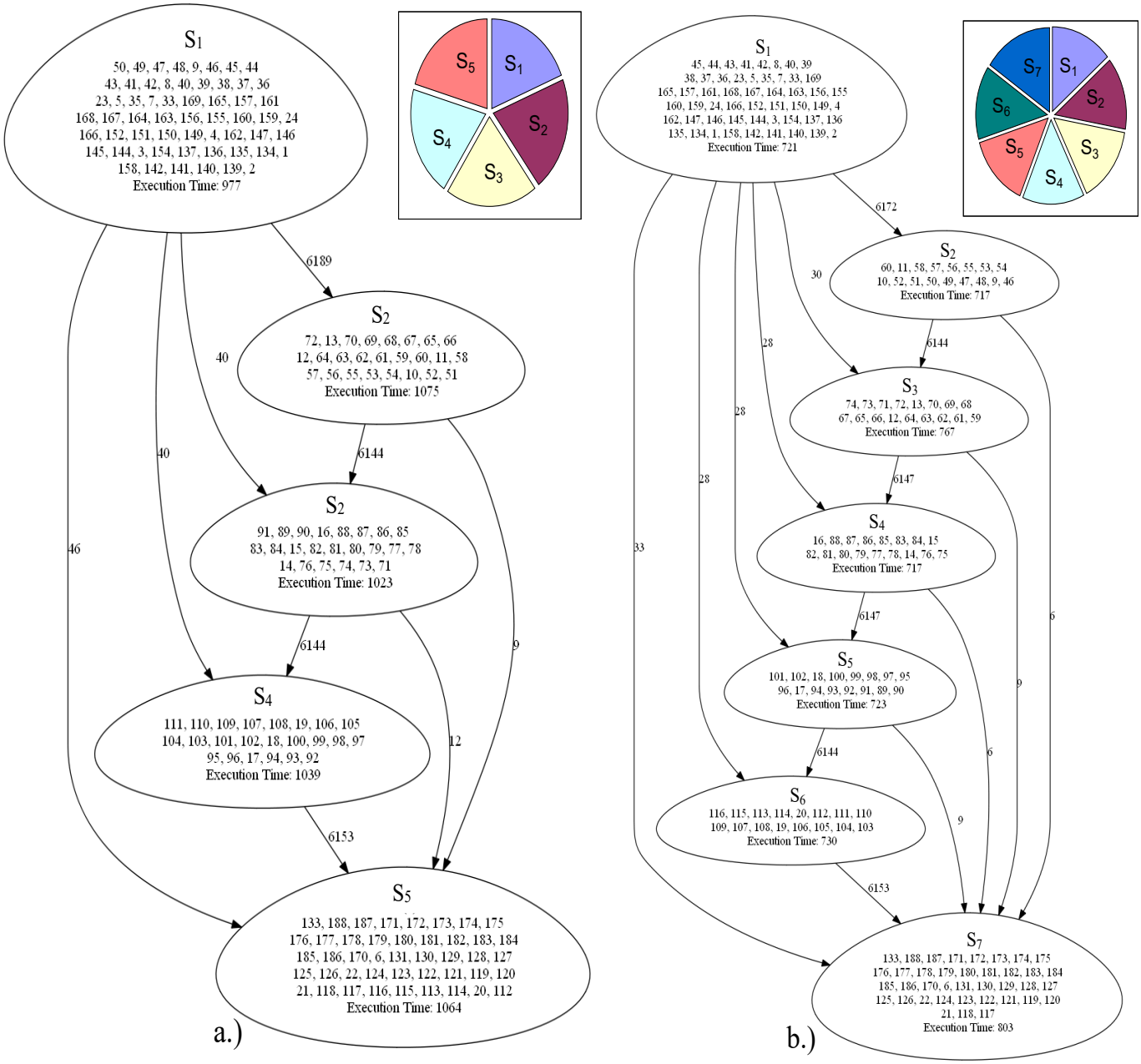


Fig. 9. Segment graphs resulted from KL decomposition (5 segments left, 7 segments right)

It is notable, that the results of different decomposition methods are almost the same. There are only minor differences in the segment contents and the execution times.

The third experiment on this benchmark is the application of DECHLS using the SC decomposition [24] algorithm without prescribing the number of segments. The SC algorithm is based on the lowest nonzero eigenvalues of the Laplacian matrix constructed from the graph [22], [30]. The lower eigenvalues we choose for the clustering, the smaller weight-sum will be caused by cut edges. Our applied version [22] does not make accessible neither the Laplacian matrix, nor its eigenvalues. Therfore, we had to calculate them by applying a separate Matlab program. Since the Laplacian matrix is a square matrix with as many rows and columns as the number of nodes in the graph, the number of eigenvalues is also the same. Depending on the complexity of the graph to be decomposed, the larger eigenvalues may be extra large. Therefore, only the smallest 25 eigenvalues of the graph representing this example are shown in Fig. 11. Since the smallest 4 nonzero eigenvalues are identical, the most beneficial decomposition could result 4 segments.

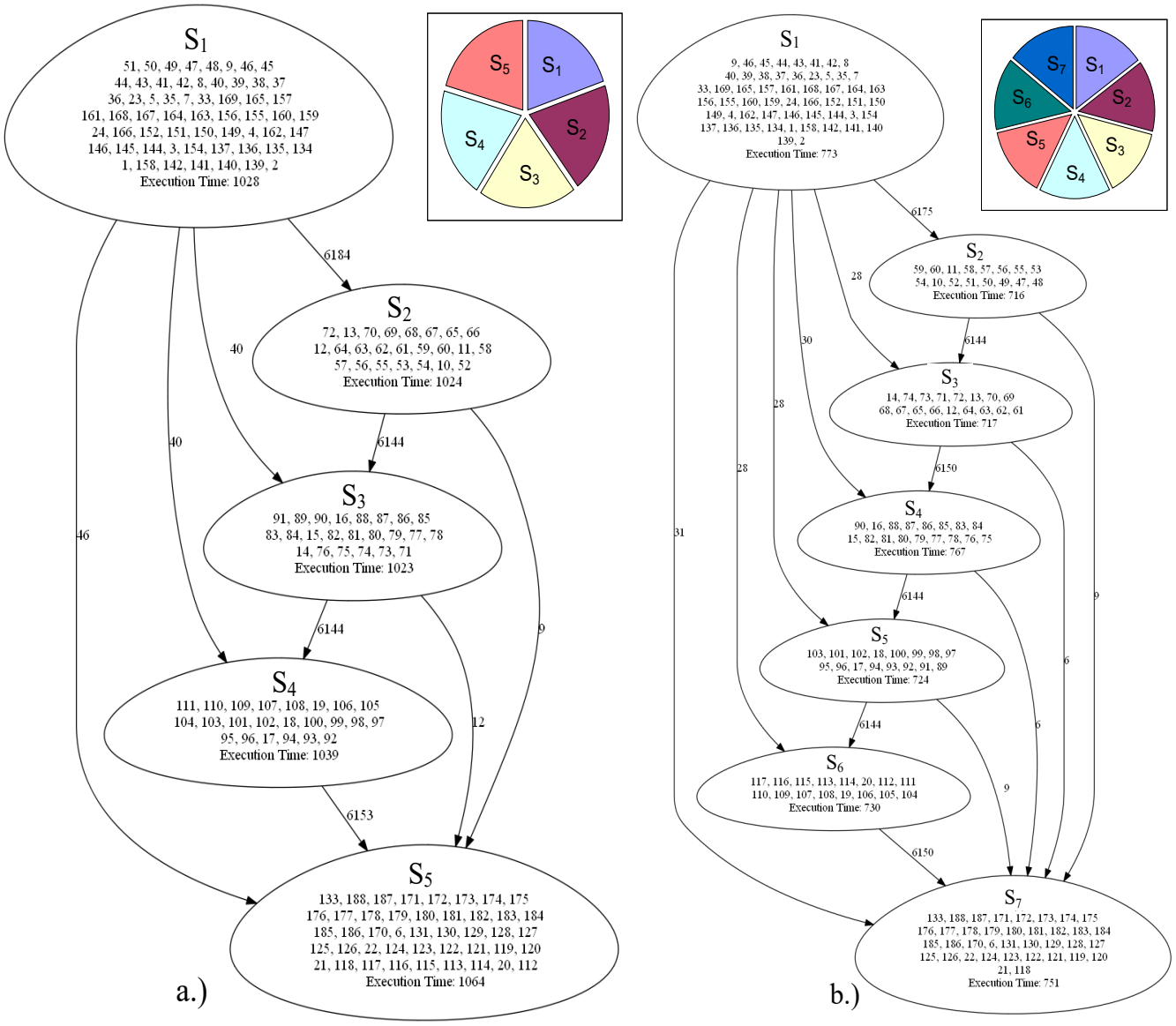


Fig. 10. Segment graphs resulted from spectral clustering decomposition (SC) (5 segments left, 7 segments right)

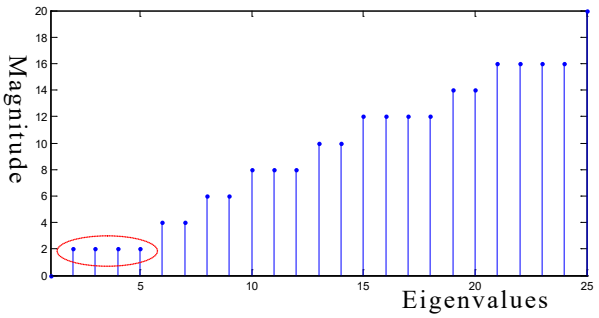


Fig. 11. Illustrating the smallest 25 eigenvalues of the Laplacian matrix used for decomposition

The result of this decomposition is shown in Fig. 12. It can be observed, that the sizes of the segments are not as uniform as in the previous cases. The reason of this is that the spectral clustering without local refinement cannot guarantee the approximately uniform segment sizes [21]. However, it is also notable, that the communication costs between the segments are now significantly lower.

Each of the resulting SGs (Figure 9, 10, 12) has been fed to our modified HLS tool in order to compute the cost versus restart time plots. Since the chosen decomposition method has a very small effect on the results, Fig. 9. a. is almost the same as Fig. 10. a. and Fig. 9. b. is almost the same as Fig. 10. b. Therefore, only three different diagrams are shown in Fig. 13.

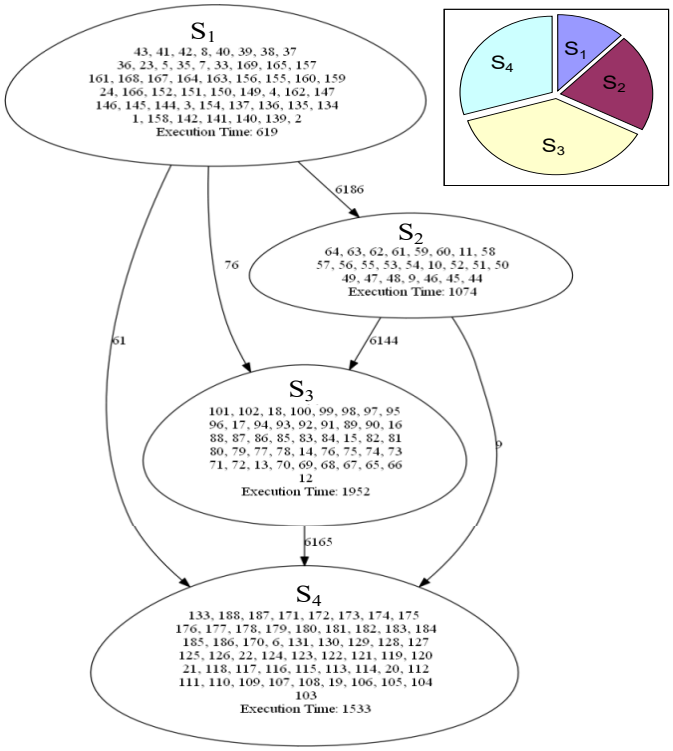


Fig. 12. Segment graph resulting from a spectral decomposition to 4 segments

4.3. *The effect of increasing the latency by DECHLS for the sound source localization*

Figures 8 and 13 illustrate that DECHLS yields significantly less processors if preliminary decomposition is included. However it can be observed on this example (Fig. 9, 10, 12) that the latency became much longer (5178) by executing DECHLS with preliminary decomposition. The latency-increasing effect of the decomposition can be observed also even in the intuitively decomposed case (Fig. 4). Usually, the decomposition may have a latency- increasing effect. One of the reasons of this could be that operations assigned to the same segment can be executed only serially. If however, the modified HLS part is executed without preliminary decomposition, then much less serial constraints arise.

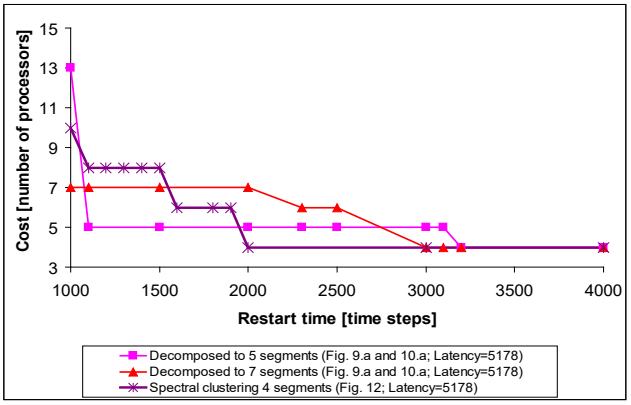


Fig. 13. Number of processors versus restart time plot after decomposition (sound source localisation)

On the other hand, a longer latency may influence advantageously the allocation and the restart time [18]. To illustrate this effect, the result of DECHLS without preliminary decomposition (Fig. 8) is recalculated (Fig. 14) at the increased latency value resulted with decomposition (Figures 9, 10, 12). In this example, further latency increasing would not have any remarkable effect, if the DECHLS is executed with preliminary decomposition as shown in Fig. 9, 10 and 12. Therefore such results are not shown in Fig. 14.

It can be observed in Fig. 14, that increasing the latency alone can provide acceptable costs, even assuming only identical processors, especially at restart time values greater than ca. 2000. The evaluation of the results in Fig. 14 would provide the most favourable solution achieved by the decomposition into 5 segments with the restart time 1100 (and the same number of processors as the number of segments in this case). However, if the restart time is less important than the cost, then the solution with 4 processors (decomposition into 4 segments by spectral clustering in this case) would be more beneficial.

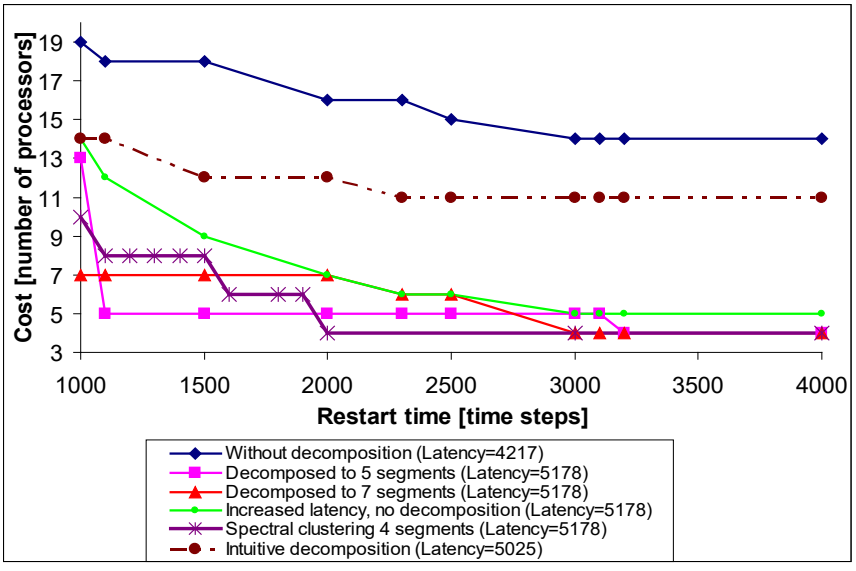


Fig. 14. Number of processors vs. restart time for comparison (sound source localisation)

4.4. A high speed industrial data logger

The second benchmark for illustration is the redesign and evaluation of a high speed data logger [31] system (Fig. 15) that performs the measurement, pre-processing and storing of several different types of input data (i.e. temperature, pressure, humidity, and some other data measured by analogue sensors). The data are to be compressed before storing them in a serial memory. The frequency of sampling and storing is the same and constant for all kinds of data. In the existing system, this task is solved by low power 8 bit microcontrollers of the PIC18FxxKxx series. The data is stored in a relatively slow, but power efficient serial storage device accessible via the I2C bus. In each input, 4 channels for pressure, temperature and humidity sensing, and 8 channels for sensing other analogue signals are assumed. A possible dataflow graph of the system is shown in Fig. 15. The required sampling data rate is at least 10 samples per second that equals to 100ms pipeline restart time (initialization interval). The design goal is to find a cost-effective solution that complies with the data rate constraint. It will be assumed that the redesigned system also applies the same types of processors and storage device.

The pre-processing subtasks of the analogue input channels are supposed to be simple moving average filters. The pressure sensing and humidity sensing subtasks are more complex. Therefore their execution times are supposed to be much longer (Table 1). The reason of this is that these values need to be temperature compensated as a part of their pre-processing.

The existing solution applies three microcontrollers as component processors. The properties of the dataflow graph allow assigning identical subtasks per microcontrollers. The main advantage of this solution is to reuse the subroutines multiple times. However, it may make the system somewhat less efficient in speed and cost. The tasks of the processors are the following. P1 performs the measurement and pre-processing the data from 3 pressure sensors and from 2 humidity sensors. P2 measures and filters the data from 8 analogue sensors and from 4 temperature sensors, because both require the same filtering. P2 also measures the data from the remaining 2 humidity sensors and from one pressure sensor. A third microcontroller (P3) is used for temperature compensation of the pressure and humidity data and also for data compression and storing tasks. The whole system is pipelined so that P1 and P2 operate in parallel forming the first pipeline stage, while P3 works on the previous set of data as the next pipeline stage. The runtimes of subtasks assigned to P1 and P2 are critical; therefore a careful adjustment should be made in order to make these runtimes as close as possible to each other. This is shown in Table 1. The microcontrollers are supposed to communicate via their SPI serial interface at the maximal available speed (4MHz). So, the data communication time-step is 0.1 ms. In order to maximize the power efficiency, the clock speed of the microcontrollers is supposed to be fixed at 4MHz, i.e. 1 ms is 10 scheduling time steps. The above assumptions (summarized in Table 1.) may yield an intuitively designable pipeline system with a restart time (initialization interval) 69 ms (690 time steps), which meets the design criteria with about 14 samples per second. The latency is 108 ms (1080 time steps) in this case.

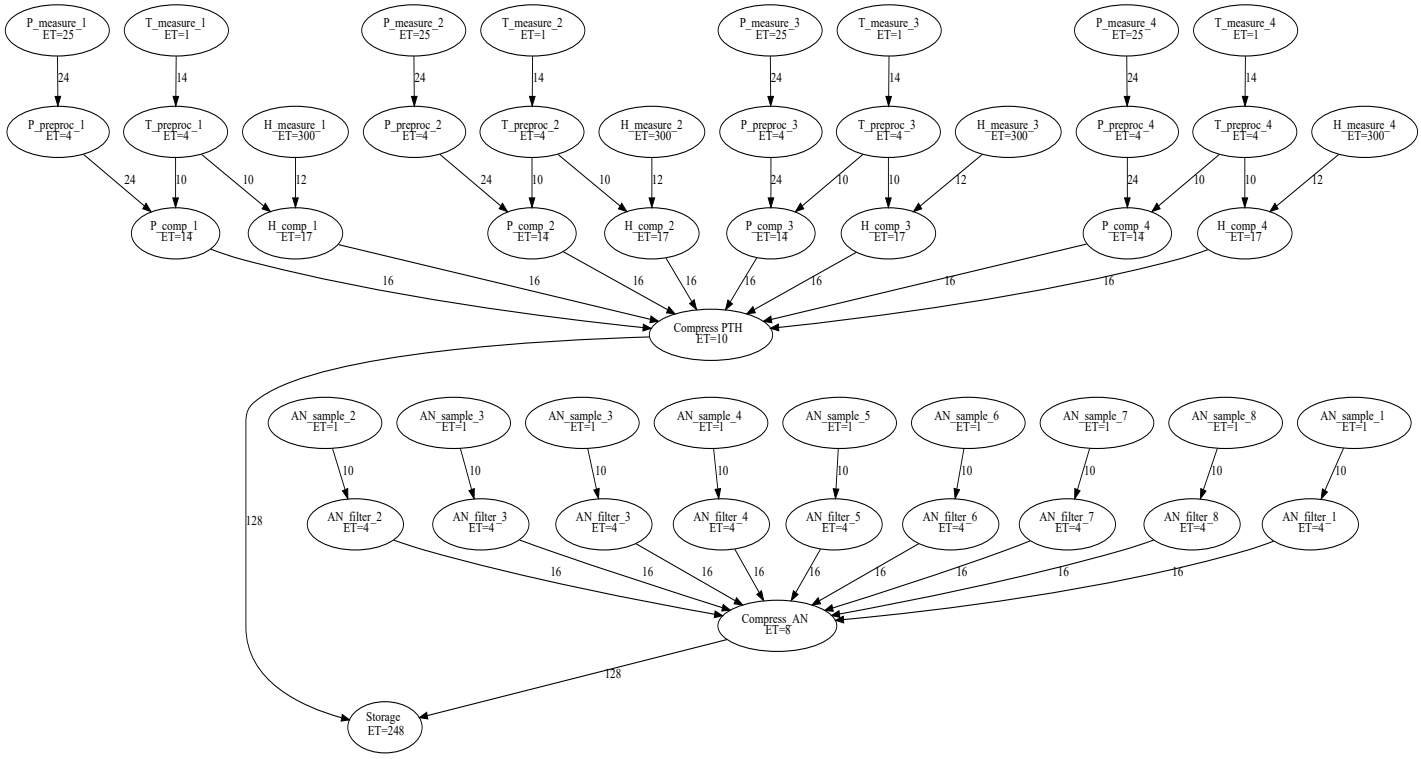


Fig. 15. Dataflow graph with execution times (time step is 0.1 ms)
(high speed data logger)

Table 1. Runtime calculation for an intuitive design

Name of task	Execution time steps	Total number	Allocated to P1	Allocated to P2	Allocated to P3
AN_sample	1	8	0	8	0
AN_filter	4	8	0	8	0
P_measurement	25	4	3	1	0
H_measurement	300	4	2	2	0
T_measurement	1	4	0	4	0
P_comp	14	4	0	0	4
H_comp	17	4	0	0	4
P_preproc	4	4	3	1	0
T_preproc	4	4	0	4	0
Compress_PTH	10	1	0	0	1
Storage	248	1	0	0	1
Compress_AN	8	1	0	0	1
Total execution time (in time steps)		1766	687	689	390

4.5. Evaluation of the existing decomposition for the high speed data logger

A carefully designed intuitive solution that complies with all the constraints was presented in Subsection 4.6. The intuitive decomposition is done firstly based on the parameter values in Table 1. The resulting segment graph is shown in Fig. 16.

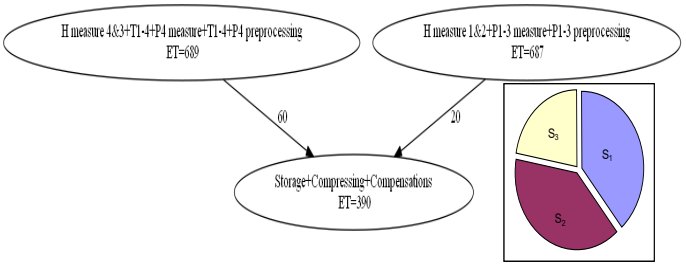


Fig. 16. Segment graph of the existing solution (high speed data logger)

In this implementation, the required number of processors is 3, the restart time is 690 time steps, and the latency is 1080 time steps. This intuitively designed system will be the basis for comparison.

4.6. Results provided by DECHLS without decomposition for the high-speed data logger

Fig. 15 shows that the dataflow graph has such a special structure in this case, which causes less than 4 processors, only if the latency is increased from 570 to above 1400 time steps. This was calculated by the modified HLS tool. If the latency is increased to 1400 time steps (140 ms), then the solution is possible by applying 3 processors. However, this case requires a longer restart time, 1050 time steps (105 ms) than the prescribed value of 1000 (100 ms) given in Subsection 4.6.

4.7. Results provided by DECHLS with preliminary decomposition for the high-speed data logger

Without prescribing the number of segments, the DECHLS yields 4 segments (Fig. 17 a). For comparison, 3, 4, 5 and 6 segments were prescribed by applying both KL and spectral clustering (SC) methods. For the sake of simplicity, only the following three cases from these are illustrated:

- 3 segments, KL method (Fig. 17. b)
- 3 segments, spectral clustering method, shortest restart time with 3 processors (Fig. 17. c)
- 5 segments, KL method, longest restart time with 3 processors (Fig. 17. d)

In this example, the number of segments and the selected decomposition method had greater influence on the results than in the sound source localisation example.

It is also notable, that the solution with 4 segments has the shortest latency, but not the longest restart time. However, the solution yielding the shortest restart time does not result in the longest latency. All of the above-mentioned cases are illustrated in Fig. 18.

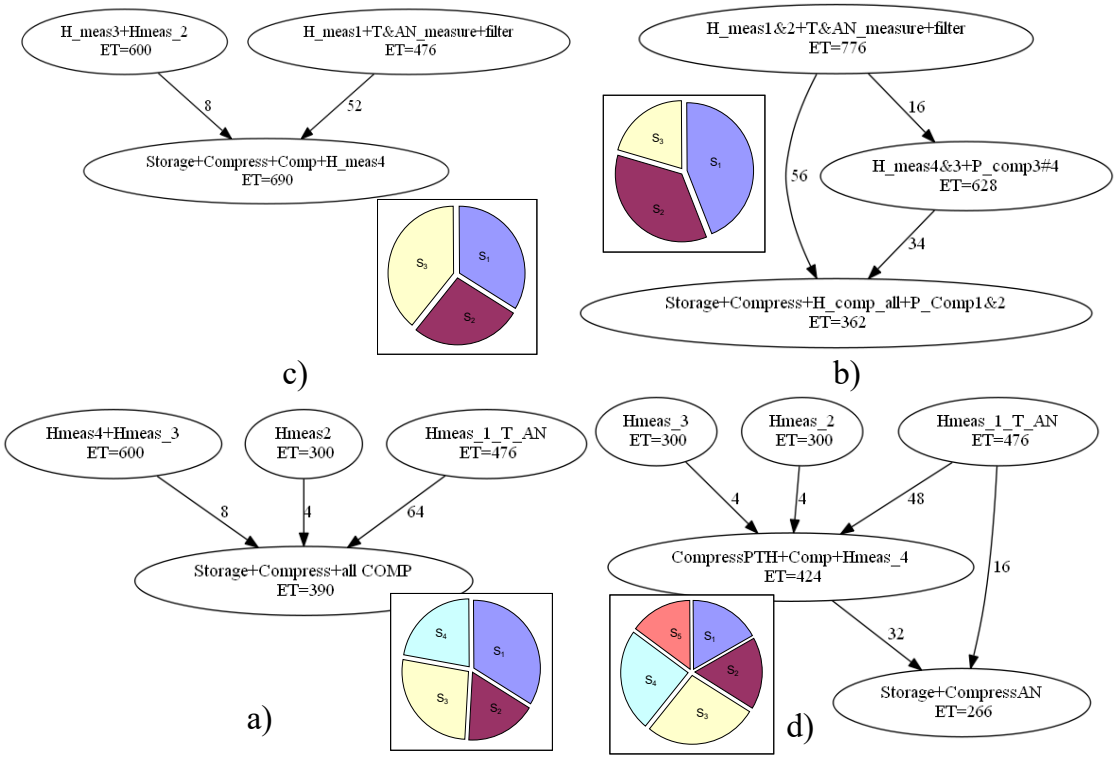


Fig. 17. Segment graphs and relative execution times obtained by some different decompositions (high speed data logger)

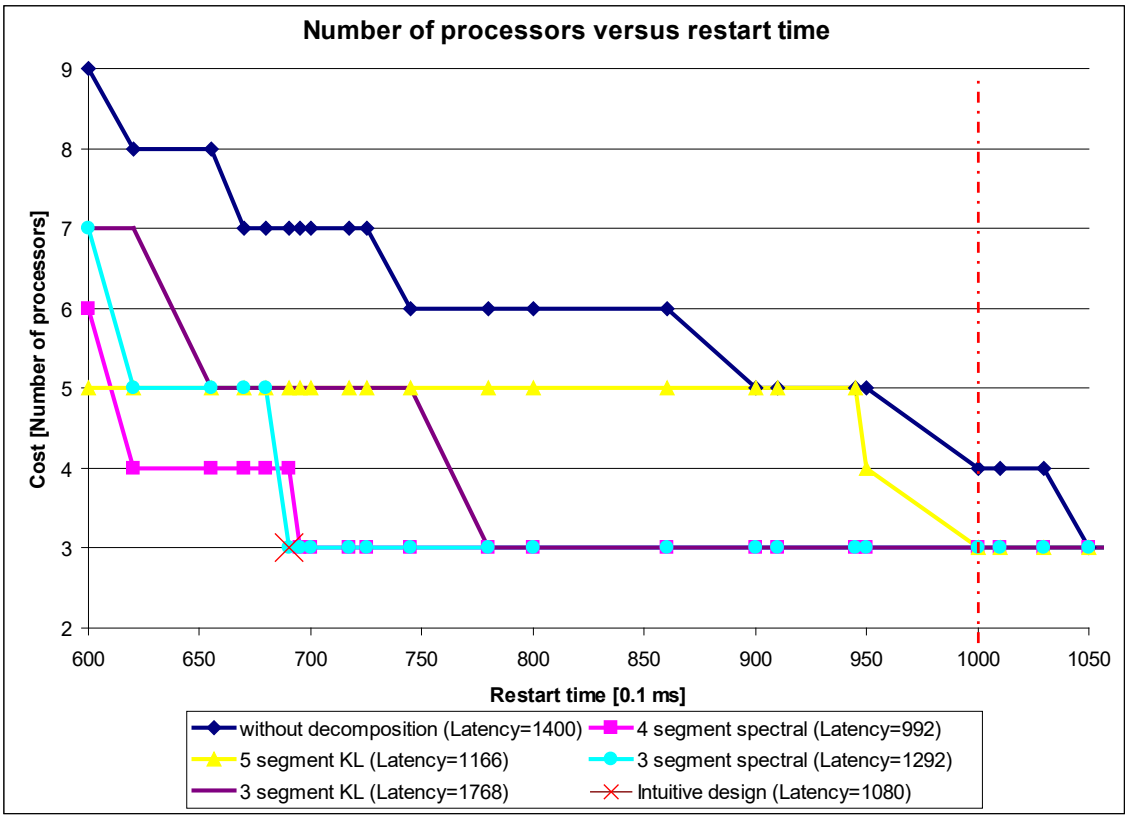


Fig. 18. Number of processors vs. restart time for comparison
(high speed data logger)

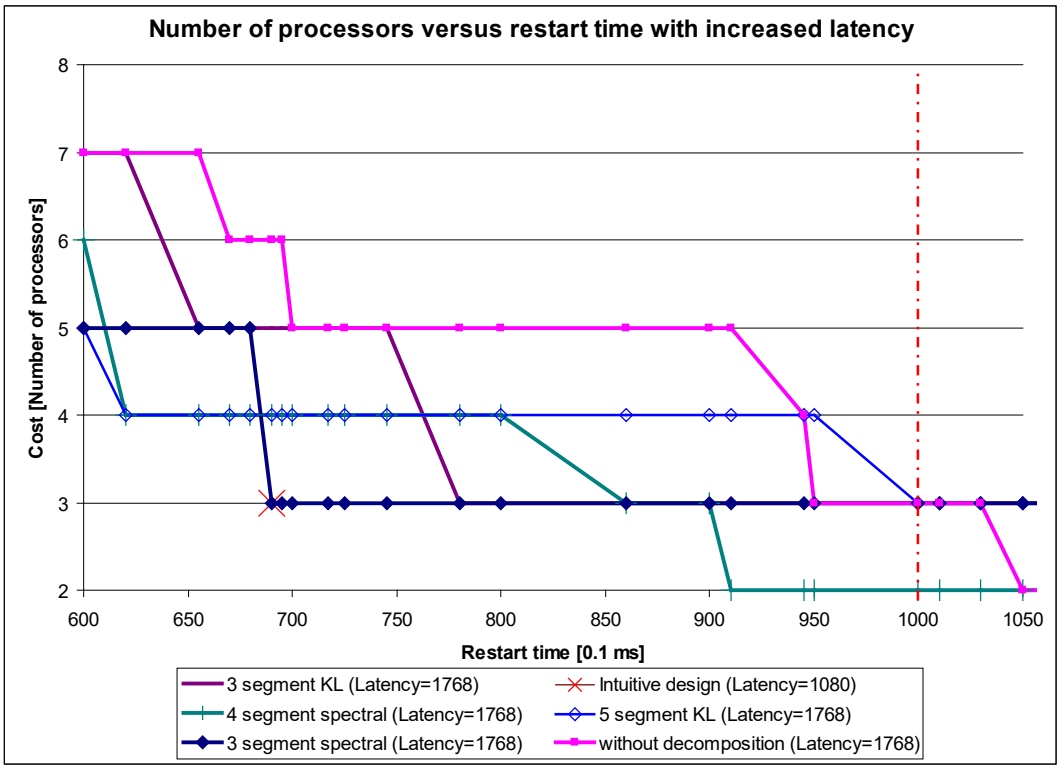


Fig. 19. Number of processors versus restart time with increased latency (high speed data logger)

4.8. The effect of increasing the latency by DECHLS for the high speed data logger

In this example, the segment graphs yield significantly different latency values. Since increasing the latency may cause a not neglectable decrease in cost, it is worthy to analyze this effect. In the above presented cases, the longest latency time obtained is L=1768. For illustration, all other cases have been recalculated after increasing the latency to this value. The results are shown in Fig. 19.

It can be observed that the increased latency provides a more cost-efficient solution by applying the spectral clustering method producing 4 segments. This result can be considered the best one with only 2 processors with pipeline restart time 920, and with the latency value 1768. Note, that the cost became lower, but the restart time and the latency time values are worse than in the intuitive implementation. However, the design constraints set in Subsection 4.6 remained still fulfilled.

5. Conclusion

The system-level synthesis phase in designing multiprocessing systems can be supported by evaluating and comparing the solutions based on various numbers of processors in respect to desired properties of the system. In this paper, a method and an experimental tool (DECHLS) have been presented for generating various system-level structures. The method and the tool DECHLS are based on the combination of preliminary task decomposition and a modified high-level synthesis tool. In its present form, this combined method applies specific decomposition algorithms and a suitably modified high-level synthesis tool as components. These components can be modified or substituted by other ones according to the purpose of the designer or to the specific properties of the target system. The application of the tool DECHLS is illustrated on redesigning two existing multiprocessing systems by analyzing and evaluating the effects on various special requirements. For both examples, several alternative solutions were obtained by DECHLS, and the most characteristic results have been plotted as the cost (number of processors) against the pipeline throughput (as restarting period or initialization interval). Based on such considerations, the designer can select the most advantageous solutions regarding the priority order of cost, pipeline throughput and latency. For the presented specific constraints, some of the obtained solutions can be considered better than the existing system. The definition and calculation of the considered parameters can be modified by the user in DECHLS.

Acknowledgment

The research work presented in this paper has been supported by the Hungarian Scientific Research Fund OTKA 72611, by the "Research University Project" TAMOP IKT T5 P3 and the research project TAMOP-4.2.2.C-11/1/KONV-2012-0004.

References

[1] Gy. Racz, P. Arato, “A Decomposition-Based System Level Synthesis Method for Heterogeneous Multiprocessor Architectures“ IEEE System On Chip Conference, IEEE 2017, Munich, Germany https://doi.org/10.1109/SOCC.2017.8226082

[2] A. Gerstlauer, C. Haubelt, A. D. Pimentel, T. P. Stefanov, D. D. Gajski, J. Teich, "Electronic System-Level Synthesis Methodologies", IEEE T COMPUT AID D., **28**(10), 1517-1530, 2009
[3] R. Corvino, A. Gamatié, M. Geilen, L. Józwiak, "Design space exploration in application-specific hardware synthesis for multiple communicating nested loops", in IEEE ICSAMOS, 128-135, 2012
[4] A. Carminati, R. S. de Oliveira, L. F. Friedrich, "Exploring the design space of multiprocessor synchronization protocols for real-time systems", Journal of Systems Architecture (JSA), **60**(3), 258-270, 2014. ISSN 1383-7621
[5] A. Cilardo, L. Gallo, N. Mazzocca, "Design space exploration for high-level synthesis of multi-threaded applications", Journal of Systems Architecture (JSA), **59**(10D), 1171-1183, 2013., ISSN 1383-7621
[6] P. Arató, Z. A. Mann, A. Orbán, "Finding optimal hardware/software partitions", Formal Methods in System Design, **31**(3), 241-263. 2007
[7] J. Wu, T. Srikanthan, G. Chen, "Algorithmic Aspects of Hardware/Software Partitioning: 1D Search Algorithms," IEEE T. COMPUT., **59**(4), 532-544, 2010. doi: 10.1109/TC.2009.173
[8] R. Ernst, J. Henkel, T. Benner, "Hardware-software cosynthesis for microcontrollers", IEEE Design & Test of Computers, **10**(4), 64-75, 1993.
[9] N. Govil, S. R. Chowdhury, "GMA: a high speed metaheuristic algorithmic approach to hardware software partitioning for Low-cost SoCs", 2015 International Symposium on Rapid System Prototyping (RSP), Amsterdam, 2015.
[10] P. Arató, T. Visegrády, I. Jankovits, High Level Synthesis of Pipelined Datapaths, John Wiley & Sons, New York, ISBN: 0 471495582 4, 2001.
[11] W. Meeus, K. Van Beeck, T. Goedemé, J. Meel, D. Stroobandt, "An overview of today's high-level synthesis tools", Design Automation for Embedded Systems, **16(**4), 31-51, 2012.
[12] T. Mukk, A. M. Rahmani, N. Dutt, "Adaptive-Reflective Middleware for Power and Energy Management in Many-Core Heterogeneous Systems", International Symposium & Workshop on Many-Core Computing: Hardware and Software, Southampton, UK, 2018.
[13] P. Arató, Z. A. Mann, A. Orbán, "Time-constrained scheduling of large pipelined datapaths", Journal of Systems Architecture (JSA), **51**(12), 665-687, 2005.
[14] P. G. Paulin, J. P. Knight, "Force-directed scheduling for the behavioral synthesis of ASICs" IEEE T COMPUT AID D., **8**(6), 661-679, 1989.
[15] M. Grajcar, "Genetic List Scheduling Algorithm for Scheduling and Allocation on a Loosely Coupled Heterogeneous Multiprocessor System", 36th ACM/IEEE Conference on Design Automation, New Orleans, LA, USA, 1999.
[16] J-H. Ding, Y-T. Chang, Z-D. Guo, K-C. Li, Y-C. Chung, "An efficient and comprehensive scheduler on Asymmetric Multicore Architecture systems", Journal of Systems Architecture, **60**(3), 305-314, 2014, ISSN 1383-7621
[17] C-S. Lin, C-S. Lin, Y-S. Lin; P-A. Hsiung, C. Shih, "Multi-objective exploitation of pipeline parallelism using clustering, replication and duplication in embedded multi-core systems", Journal of Systems Architecture (JSA), **59**(10C), 1083-1094, 2013, ISSN 1383-7621
[18] Gy. Pilászy, Gy. Rácz, P. Arató, "The effect of increasing the latency time in High Level Synthesis", Periodica Polytechnica Electrical Engineering, **58**(2), 37-42, 2014. https://doi.org/10.3311/PPee.7024
[19] K. O'Neal, D. Grissom, P. Brisk, "Force-Directed List Scheduling for Digital Microfluidic Biochips", IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC), Santa Cruz, USA, 2012.
[20] P. G. Paulin, J. P. Knight, "Force-Directed Scheduling in Automatic Data Path Synthesis", 24th Conference on Design Automation, New York, USA, 1987.
[21] P. Arató, D. Drexler, Gy. Rácz, "Analyzing the Effect of Decomposition Algorithms on the Heterogeneous Multiprocessing Architectures in System Level Synthesis" Buletinul Stiintific al Universitatii Politehnica din Timisoara Romania Seria Automatica si Calculatore / Scientific Buletin of Politechnica University of Timisoara Transactions on Automatic Control and Computer Science, **60**(74) 39-46. 2015.
[22] B. Hendrickson, R. Leland, "The Chaco User's Guide Version 2.0", Technical Report SAND94-2692, 1994.
[23] B. Hendrickson, R. Leland, "A Multilevel Algorithm for Partitioning Graphs", In Proc. Supercomputing '95. (Formerly, Technical Report SAND93-1301), 1993.
[24] B. Hendrickson, R. Leland, "An Improved Spectral Graph Partitioning Algorithm for Mapping Parallel Computations", SIAM J. Sci. Stat. Comput., **16**(2), 452-469, 1995.
[25] G. Karypis, V. Kumar, "Metis-unstructured graph partitioning and sparse matrix ordering system version 2.0", 1995.
[26] A. Trifunovic, W. J. Knottenbelt, "Parkway 2.0: a parallel multilevel hypergraph partitioning tool", Computer and Information Sciences-ISCIS 2004, Springer, 789-800. 2004.
[27] M. Goraczko, J. Liu, D. Lymberopoulos, "Energy-Optimal Software Partitioning in Heterogeneous Multiprocessor Embedded Systems", DAC 2008., Anaheim, California, USA, 2008.
[28] G. Suba, "Hierarchical pipelining of nested loops in high-level synthesis", Periodica Polytechnica Electrical Engineering, **58**(3) 81-90, 2014. https://doi.org/10.3311/PPee.7610
[29] J. Subhlok, G. Vondran, "Optimal latency - throughput tradeoffs for data parallel pipelines", 8th Annual ACM Symposium on Parallel Algorithms and Architectures, Padua, Italy, 1996.
[30] C. J. Alpert, A. B. Kahng, S.-Z. Yao, "Spectral partitioning with multiple eigenvectors", Discrete Applied Mathematics, **90**(1), 3-26, 1999.
[31] Project documentation of KMR_12-1-2012-0222 titled "The development of a complex monitoring system for fishing waters"

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 130-172 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Acoustic Signal Processing and Noise Characterization Theory via Energy Conversion in a PV Solar Wall Device with Ventilation through a Room

Himanshu Dehra*

Egis India Consulting Engineers Pvt. Ltd., Egis Tower, Plot # 66, Sector 32, Gurugram (Haryana), 122001, India

ARTICLEINFO

Article history:
Received: 04 June, 2018
Accepted: 24 July, 2018
Online: 29 July, 2018

Keywords:
Acoustics
Signal Processing
Noise Characterization
Energy Conversion
Ventilation
Loudspeaker
PV Solar Wall Device

ABSTRACT

Noise defined as 'a sensation of unwanted intensity of a wave', is perception of a pollutant and a type of environmental stressor. The unwanted intensity of a wave is a propagation of noise due to transmission of waves (viz. physical agents) such as light, sound, heat, electricity, fluid and fire. The characterization of noise interference, due to power difference of two intensities in a wave is presented. Noise interference characterization in a wave is obtained depending on type of wave. Standard definitions of noise sources, their measurement equations, their units and their origins under limiting reference conditions are derived. All types of wave form one positive power cycle and one negative power cycle. The positive and negative noise scales and their units are devised depending on speed of noise interference in a wave. A numerical and experimental study was conducted for supporting the noise characterization theory via ascertainment of energy conversion characteristics of a pair of photovoltaic (PV) modules integrated with solar wall of an outdoor test-room. A pre-fabricated outdoor room was setup for conducting outdoor experiments on a PV solar wall with ventilation through the outdoor room. Acoustic signal processing is supported with some experimental and numerical results of a parallel plate PV solar wall device installed in an outdoor test-room to authenticate the noise interference equations. Detailed discussions on noise characterization theory along with some examples of noise filter systems as per noise sources are also presented. The noise characterization theory is also exemplified with some noise unit calculations using presented noise measurement equations.

1. Introduction

The aim of this paper is to present acoustic signal processing and noise characterization theory with support of energy conversion models and experiments on a photovoltaic (PV) solar wall with ventilation through a room. This paper is an extension of work originally presented in conference [1, 2].

Noise, defined as 'a sensation of unwanted intensity of a wave', is perception of a pollutant and a type of environmental stressor. An environmental stressor such as noise may have detrimental effects on various aspects of health. The unwanted intensity of a wave is a propagation of noise due to transmission of waves (viz. physical agents) such as light, sound, heat, electricity, fluid and fire [1]. A unified theory for stresses and oscillations is applicable so as to take into effect of all the physical agents as an environmental stressor on a human body [3]. As per the theory, the stresses acted on a particle due to interaction of many forces are distinguished as fundamental, internal and external stresses. The existence of fundamental stresses is due to presence of electromagnetic and gravitational forces. The internal stresses are acted under the influence of fundamental stresses and are constituted by composition and properties of a particle. The external stresses are acted under the existence of an external source of energy.

Solar energy conversion occurs at solar cells and solar intensity of incident solar energy is converted into electric power and waste heat. The photovoltaic devices with ventilation provide means for converting waste heat lost to surrounding environment into useful thermal power [2-7]. The composite waves are transmitted in photovoltaic (PV) devices due to stresses and oscillations of incident solar and ventilation energy. In this way solar power intensity is converted into heat, fluid, electricity, light, sound and fire depending on intensities

* Corresponding author: Himanshu Dehra, Email: anshu_dehra@hotmail.com

of its transmitted composite waves in a PV solar wall device.

1.1. Methodology

This paper has summarized the concept of acoustic signal processing and noise characterization theory with aid of energy conversion in a PV solar wall device. Basic definitions on noise characterization are presented in Appendix. Sources of noise are characterized as per physiological responses from human senses. Definitions of various types of noise as per type and speed of wave are presented. Noise measurement equations are deduced from the wave theory. Acoustic or noise filters as per source of various power systems are defined and illustrated with examples. Experimental results on energy conversion are presented on a PV solar wall device. Sensible heat storage capacity and thermal storage capacities are explained for their energy conversion phenomenon.

Some numerical results are also presented to support the noise characterization. In order to further authenticate the noise wave characterization, signal processing is achieved from a PV device composed of RC analogue signal. Acoustic signal processing is supported with some experimental and numerical results of a parallel plate PV solar wall device installed in an outdoor test-room for verification of the noise interference equations. Some noise unit calculations deduced from the numerical model based on newly devised noise measurement equations are also presented in tables. Detailed discussions are also presented on significance of noise characterization theory for: i) cities; ii) acoustics; iii) electrical engineering, instrumentation & control; iv) energy conversion in engines; v) noise pollution and control; vi) fire alarm systems; vii) operation of power systems; viii) friction & oscillations; ix) noise instrumentation; x) automatic control & management in noise systems; xi) potential research themes and graduate teaching subjects.

The following sections define and describe 'noise', its sources, its measurement equations with support of experimental and numerical results of a PV solar wall device with ventilation through a room. Discussions of noise characterization theory on solar energy acoustics, energy conversion model and experiments, acoustic signal processing, noise filters & their examples with description of sensors & transducers for a human brain along with human comfort & health are presented in later sections.

2. Noise Characterization

2.1. Source & Sink of Noise

The strength of a source is defined as total energy flow per unit time and unit length of line. A line which is drawn hypothetically perpendicular to the plane, from which energy is assumed to flow uniformly in all directions at normal angles to it, is defined as a source. It acts as a point in the usual two-dimensional energy flow diagram. The flow is in imaginary radial lines from the source, the current of energy flow is at a distance r from the source. The current is determined by the strength divided by the energy flow area [8-10].

2.2. Psychoacoustics in New Dimensions

The standard sources of noise along with their definitions, their measurement equations and measurement units are presented. A contemporary theory of psychoacoustics was presented for characterization of interference of noise waves due to difference of power of two intensities [11]. The difference of two intensities of power is due to transmission of heat, fluid, light, electricity, fire and sound into a particle body. The sources of noise are classified according to the type of wave of interference. The various types of noise are defined with factors of speed of wave, areas of energy stored in a wave, due to interference and difference of power between two intensities of wave. Noise filters as per source of noise signals from noise power systems are differentiated for filtering unwanted frequencies from solar power, electric power, light power, sound power, heat power, fluid power and fire power.

The interference of noise arises due to difference of power of two intensities [1-13]. The intensity of power for any particle body is a function of development of various stresses. The phenomenon of acoustic resonance arises due to matching of critical stress level with the natural stress level necessary for oscillation of a particle body. The criteria for generation of acoustic resonance include waves propagated with transmission of light, sound, noise, heat, electricity, fluid and fire from a particle body. The psychological feeling of sensation and perception of noise from light, sound, heat, electricity, fluid and fire is a physiological response from the sensory organs of a standard (average) human body.

3. Sources of Noise

The sources of noise are classified according to the type of wave of interference:

Light: In the electromagnetic radiation wavelength band from approximately between 380-765 nm, the visual sensation of light is tested by the eye of an observer with seeing of a radiant energy. The physiological response from an average eye defines units of light. The sensitivity of human eye is not same in all wavelengths or colors. The contribution of adding daylight is visual sensation in the visible region of the solar energy spectrum.

Sound: In the range of frequencies between 20 Hz and 20,000 Hz, the sound is evaluated due to presence of fluid pressure energy as a hearing sensation by the ear. The sound units are based on the functional feedback of an average ear. The sensitivity of sound to the whole frequency band is not same for human ear.

Heat: In the electromagnetic radiation between 0.1 μm to 100 μm, the heat as a temperature sensation is examined by the human body. The sensation function of temperature defines units of heat. The temperature sensation function is a measure of coldness and hotness. The comfort zone of temperature is evaluated from functional feedback of a human body which also defines thermal comfort. The contribution to discomfort of human body is in the ultra violet region of solar energy spectrum.

Electricity: With passing of direct current or an alternating

current, the electricity as a shock sensation is evaluated by skin of an observer due to electromagnetic energy stored in a conductor which is short-circuited by a human body.

Fluid: The fluid as combined ventilation and breathing sensation is evaluated by amount of fluid passed either externally or internally through a standard (average) human body.

Fire: The exposure of radiant energy and fluid acting on skin surface of an average human body, defines the fire as a sensation of burning.

4. Noise Measurement Equations

The following standard measurement equations are derived and adopted from the standard definitions for sources of noise interference as mentioned in Appendix.

Noise of Sol: For a pack of solar energy wave, the multiplication of solar power storage and the velocity of light gives solar power intensity I. On taking logarithm of two intensities of solar power, I_1 and I_2, provides intensity difference. It is mathematically expressed as:

$$Sol = \log(I_1)(I_2)^{-1} \quad (1)$$

Whereas logarithmic unit ratio for noise of sol is expressed as *Sol*. The oncisol (oS) is more convenient for solar power systems. The mathematical expression by the following equality gives an oncisol (oS), which is $1/11^{th}$ unit of a *Sol*:

$$oS = \pm 11\log(I_1)(I_2)^{-1} \quad (2)$$

Noise of Therm: For a pack of heat energy wave, the multiplication of total power storage and the velocity of light gives heat power intensity I. The pack of solar energy wave and heat energy wave (for same intensity I), have same energy areas, therefore their units of noise are same as *Sol*.

Noise of Photons: For a pack of light energy beam, the multiplication of total power storage and the velocity of light gives light power intensity I. The pack of solar energy wave and light energy beam (for same intensity I), have same energy areas, therefore their units of noise are same as *Sol*.

Noise of Electrons: For a pack of electricity wave, the multiplication of total electrical storage and the velocity of light gives electrical power intensity I. The pack of solar energy wave and electricity wave (for same intensity I), have same energy areas, therefore their units of noise are same as *Sol*.

Noise of Scattering: For a pack of fluid energy wave, the multiplication of total power storage and the velocity of fluid gives fluid power intensity I. On taking logarithm of two intensities of fluid power, I_1 and I_2, provides intensity difference. It is mathematically expressed as:

$$Sip = \log(I_1)(I_2)^{-1} \quad (3)$$

Whereas, logarithmic unit ratio for noise of scattering is *Sip*. The oncisip (oS) is more convenient for fluid power systems.

The mathematical expression by the following equality gives an oncisip (oS), which is $1/11^{th}$ unit of a *Sip*:

$$oS = \pm 11\log(I_1)(I_2)^{-1} \quad (4)$$

For energy area determination of a fluid wave, the water with a specific gravity of 1.0 is the standard fluid considered with power of ±1 Wm^{-2} for reference intensity I_2.

Noise of Scattering and Lightning: For a pack of fire wave, the intensity I, of fire flash with power of light, is the multiplication of total power storage and the velocity of light. Whereas for a pack of fire wave, the intensity, I, of fire flash with power of fluid, is the multiplication of total power storage capacity and velocity of fluid.

- For a noise due to fire flash, the collective effect of scattering and lightning is to be obtained by superimposition principle.
- For same intensity I, the pack of solar energy wave and a fire flash with light power have same energy areas, therefore their units of noise are same as *Sol*. The therm power may also be included in fire flash with power of light.
- For same intensity I, the pack of fluid energy wave and a fire flash with fluid power have same energy areas, therefore their units of noise are same as *Sip*. In determining the areas of energy for the case of fluids other than water, a multiplication factor in specific gravity of fluid is to be considered.

Noise of Elasticity: For a pack of sound energy wave, the product of total power storage and the velocity of sound gives sound power intensity I. On taking logarithm of two intensities of sound power, I_1 and I_2, provides intensity difference. It is mathematically expressed as:

$$Bel = \log(I_1)(I_2)^{-1} \quad (5)$$

Whereas, logarithmic unit ratio for noise of elasticity is *Bel*. The oncibel (oB) is more convenient for sound power systems. The mathematical expression by the following equality gives an oncibel (oB), which is $1/11^{th}$ unit of a *Bel*:

$$oB = \pm 11\log(I_1)(I_2)^{-1} \quad (6)$$

4.1. Limiting Conditions

The Table 1 summarises units of noise and their limiting conditions.

Table 1: Noise under Limiting Conditions

Reference[a] (I_2 = ±1 Wm^{-2})	Noise Scales and limiting Conditions		
	Noise of Sol	**Noise of Scattering**	**Noise of Elasticity**
Units	**Sol**	**Sip**	**Bel**
I_1 = 1 Wm^{-2}	No Positive Solar Energy	No Positive Fluid Energy	No Positive Sound Energy
I_1 = 1+→0 Wm^{-2}	Decreasing Solar Energy	Decreasing Fluid Energy	Decreasing Sound Energy
I_1= + ve	Increasing Solar Energy	Increasing Fluid Energy	Increasing Sound Energy
I_1 = -1 Wm^{-2}	Negative Solar Energy	Negative Fluid Energy	Negative Sound Energy
	Darkness	Low Pressure	Inaudible range
I_1 = -ve	Darkness increasing, distance from point source of light increasing	Low pressure increasing, vacuum approaching	Inaudible range increasing, vacuum approaching
I_1 = -1+→0 Wm^{-2}	Negative Solar Energy	Negative Fluid Energy	Negative Sound Energy
	Decreasing Darkness	Decreasing Low Pressure	Decreasing inaudible range
* Reference value of I_2 = ±1 Wm^{-2} signifies the limiting condition with areas of noise interference approaching to zero.			

There are following elaborative points on choosing an *onci* as $1/11^{th}$ unit of noise [1]:

i) Reference value used for I_2 is -1 W m^{-2} on positive scale of

noise and 1 W m^{-2} on negative scale of noise. In a power cycle, all types of wave form one positive power cycle and one negative power cycle (see Figure 10). Positive scale of noise has 10 positive units and one negative unit. Whereas, negative scale of noise has 1 positive unit and 10 negative units;

ii) Each unit of sol, sip and bel is divided into 11 parts, 1 part is $1/11^{th}$ unit of noise;

iii) The base of logarithm used in noise measurement equations is 11;

iv) Reference value of I_2 is -1 W m^{-2} with I_1 on positive scale of noise, should be taken with negative noise measurement expression (see Eqs 2, 4 and 6), therefore it gives positive values of noise;

v) Reference value of I_2 is 1 W m^{-2} with I_1 on negative scale of noise, should be taken with positive noise measurement expression (see Eqs 2, 4 and 6), therefore it gives negative values of noise.

5. Noise Filters

The noise filters are classified as per source signal of unwanted frequencies from solar power, electric power, light power, sound power, heat power, fluid power and fire power. An acoustic filter is an electrical analog circuit of various combinations of RC feedback circuit with an operational amplifier and is used to filter unwanted frequencies of oscillations from a power system [14]. It is a network with selective transmission for currents from a power system of varying frequency. The unwanted frequencies generated from a power system are removed by using an operational amplifier with different combination of filter arrangements. An operational amplifier is an integrated circuit that consists of several bipolar transistors, resistors, diodes, and capacitors, interconnected so that amplification can be achieved over a wide range of frequencies [15].

The action of filtering the frequency from a power system is based on the variation in the reactance of an inductance or a capacitance with change in frequency. The band of frequencies that can be removed from a power system can be at the low frequency end of frequency spectrum, at the high frequency end, at both ends, or in the middle of the spectrum. The filters to perform each of these operations are known respectively as low-pass filters, high-pass filters, band-pass filters and band-stop filters. There are many configurations of design of filters. The filters are divided into passive and active configurations. The passive filters are less effective simple circuits constructed with resistors, capacitors, and inductors. The active filters are useful in providing an effective filtering action than passive filters. The active filters require a source of operating power.

5.1. Noise Filter Systems

The criteria for definitions of filters for noise filtering is based on areas of energy stored in a wave due to noise interference, speed of wave and difference of power between two intensities of wave. The filtered noise signals are considered from systems of solar power, electric power, light power, sound power, heat power, fluid power and fire power. The noise filters as per sources of noise are defined as follows:

Filter for noise of sol: This filter is used to filter noise due to difference of power intensities between two solar power systems. Example: window curtain, window blind, wall and glass.

Filter for noise of therm: This filter is used to filter noise due to difference of power intensities between two heat power systems. Example: house, insulation, clothing and furnace.

Filter for noise of photons: This filter is used to filter noise due to difference of power intensities between two light power systems. Example: 3-D vision of any object, electric bulb, television, computer and LCD screen laptop.

Filter for noise of electrons: This filter is used to filter noise due to difference of power intensities between two electrical power systems. Example: AM/FM radio clock with ear phones, telephone instrument with ear phones and CD audio player with ear phones.

Filter for noise of scattering: This filter is used to filter noise due to difference of power intensities between two fluid power systems. Example: electric fan, pump, motor vehicle, river stream and tap water.

Filter for noise of scattering and lightning: This filter is used to filter noise due to difference of power intensities between two fire power systems. Example: lighter, matchstick, gas stove, locomotive engine and thunder-bolt.

Filter for noise of elasticity: This filter is used to filter noise due to difference of power intensities between two sound power systems. Example: your vocal apparatus, organ pipe, thunder-bolt and drum beats.

6. Some Examples of Noise Filters

Some examples of noise filters are enumerated as under.

6.1. Human Voice Production

The example of phonetics of filtering sound of a human speech is illustrated. The human speech is synthesized due to development of stresses at vocal folds [16-17]. The smoothening of the sound is function of its amplitude and its shape of oscillations at vocal tract of a human being. The vocal tract is a resonant cavity wall with sound energy stored in oscillations of its vocal folds. The vocal chords showing mechanism of synthesis of human speech is illustrated in Figure 1.

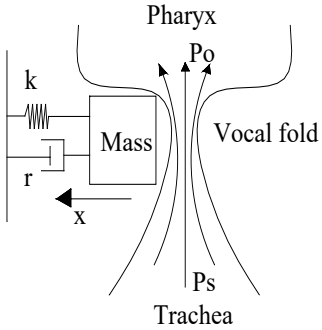


Figure 1: A human vocal mechanism

6.2. An Airflow Window with a Photovoltaic Solar Wall

The filtering of solar energy is illustrated through an example of an airflow window attached with a shading device. An airflow window is fixed with a movable roller blind to control the transmission of daylight as well as amount of solar heat. The bottom portion of photovoltaic solar wall is used for controlling the amount of air ventilation along with generation of solar electric power. The example is illustrated in Figure 2.

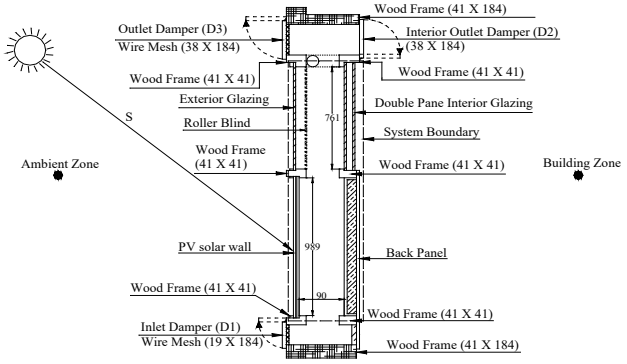


Figure 2: An Airflow window with a Photovoltaic Solar Wall (Dimensions shown are in mm).

6.3. Psychrometric Air Conditioner

An elementary air conditioner for summer comfort conditioning consists of a cooling coil, a cooling fluid with a filter [18]. The schematic of operation of a psychrometric air conditioner is illustrated in Figure 3.

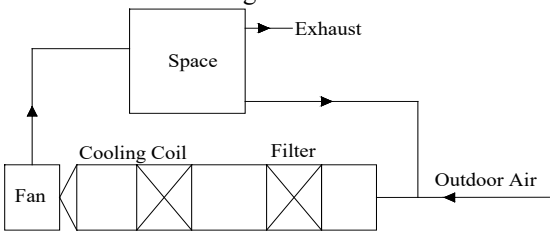


Figure 3. A psychrometric air conditioner

6.4. Telephone Line

The impedance of a telephone line is composed of distributed resistance, capacitance, and inductance [19]. The impedance of telephone line is proportional to the insulation, loop length and whether the wire is buried, aerial or bare parallel wires strung on telephone pole. A telephone line is usually supplied with a 48 VDC from the telephone exchange. The schematic of operation of a telephone line with telephone instrument is illustrated in Figure 4.

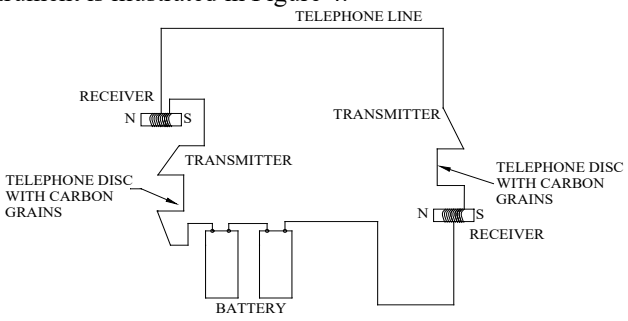


Figure 4: Operation of a Telephone Line.

6.5. Fire and Smoke Detection System

A fire detection system consists of a control system with interconnected alarms, smoke and heat detectors. A fire detector is a device which is used for presetting an alarm at a particular temperature. A smoke detector is a device which is used for presetting an alarm when a certain percentage of smoke accumulates. The photovoltaic cell activates the smoke alarm only if it senses requisite obscuration of light over a unit area with control from BMS [20]. The schematic of various components for fire detection system is illustrated in Figure 5. The detail description of a fire alarm systems is provided later in discussions.

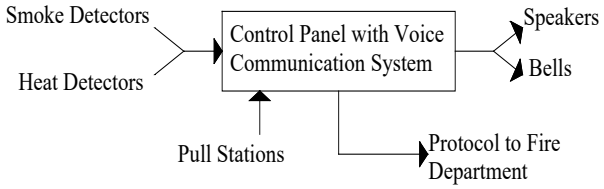


Figure 5: A fire detection system

7. Experiments

Many studies on facade integrated PV system are focused on the model simulation of radiation and convection [21-36]. Kim and Kim [28] compared three different types of integrated PV-thermal system configurations. The study analyzed the system performance as well as the building performance. Various researchers [29-30] have also conducted research on the energy performance and efficiency of actively air cooled PV modules integrated into facades.

In the present study, the salient feature of a pair of PV modules used in solar wall was the glass coating surrounding the semiconductor material of solar cells [24-25]. The manufacturing of glass coated PV modules was based on multi-crystalline solar cell technology. The operating conditions were selected for conducting various experiments on a PV solar wall. The indoor solar simulator tests were conducted for PV modules under standard operating conditions. The outdoor tests were conducted under available climatic conditions for obtaining temperatures, air velocities, thermal time constants, heat capacity, and thermal storage capacity of a PV solar wall. The PV modules used for experiments were having a glass coating of 3 mm attached on its exterior and interior sides. The glass coating was attached to exterior and interior sides of PV modules in the form of glass sheets by a double-glass (PWX) manufacturing technology. Each PV module was having 36 multi-crystalline solar cell units, with thin transparent gaps in between them. The glass coating on PV modules has improved the overall performance of PV modules in many ways; it has: i) protected semi-conductor material from weather deterioration; ii) absorbed ultra violet rays of solar energy and trap solar heat in glass coating, thus increasing the temperature of the surrounding environment; iii) reduced the reflection losses of sunlight rays reflected from PV module; iv) improved the aesthetic appearance of PV module; v) increased the thermal

mass of PV module; and vi) enhanced the structural strength of PV module. The pictorial view of glass coated PV module used in experiments is illustrated in Figure 6. The typical characteristics of glass coated PV module without frame were: dimensions of 993 mm by 453 mm and a thickness of 6.2 mm with glass coating, typical current of 2.8 A, typical voltage of 17 V, temperature coefficients of current and voltage were + 0.034 %/°C and -2.17 mV/°C/solar cell at standard test conditions [2, 24].

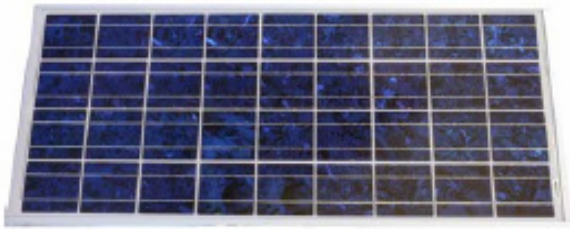


Figure 6: Glass coated PV module used in experiments

The indoor measurement tests were carried out in an indoor simulation test facility of CANMET Energy Technology Centre of Natural Resources, Varennes, Québec, Canada [2]. The schematic of indoor solar simulator used for establishing characteristics of glass coated PV modules is illustrated in Figure 7.

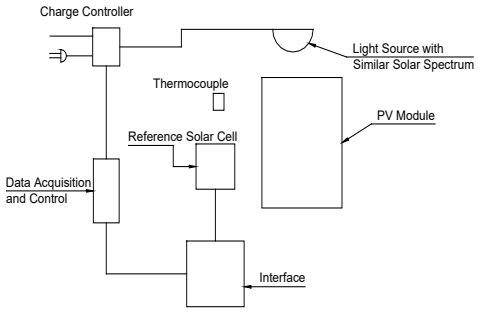


Figure 7 Block diagram of an indoor solar simulator

The current-voltage and power-voltage characteristics of one of the PV modules are shown in Figs. 8 and 9 [2]. Table 2 has presented characteristics of a pair of PV modules established with an indoor solar simulator [2]. As presented in Table 2, electrical conversion efficiency to the order of 8.9% was reported from indoor simulator tests. The current-voltage and power-voltage characteristics of tested PV module are illustrated in Figs. 8 and 9.

Table 2: Characteristics of glass coated pv modules established with an indoor solar simulator

PV Module	Solar Intensity (Wm^{-2})	Spectrum	Isc (A)	Voc (V)
PW30	1009.11	1.5 A.M. Class B	2.833	21.296
PW31	1007.58	1.5 A.M. Class B	2.829	21.352
PV Module	Vmax (V)	Pmax (Watts)	Fill Factor	Efficiency (%)
PW30	16.198	42.157	0.6985	8.703
PW31	16.503	43.435	0.7188	8.981

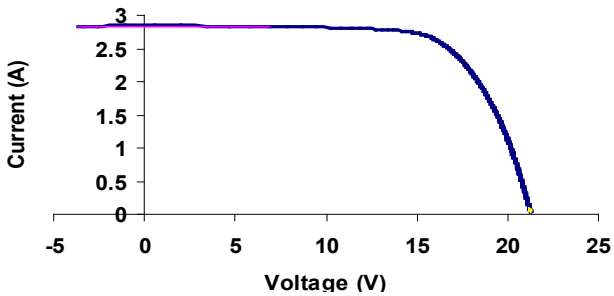


Figure 8: Current-voltage characteristics of PW30

After determining current-voltage and power-voltage characteristics from an indoor solar simulator, the pair of glass coated PV modules was installed in the test section of an outdoor test-room. The outdoor experiments were conducted on PV modules in order to determine energy conversion characteristics and further improving the overall performance of PV modules by using thermal mass of an outdoor room.

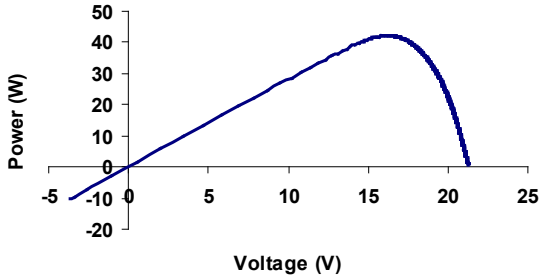


Figure 9: Power-voltage characteristics of PW30

The outdoor room was a pre-fabricated structure that was delivered into the premises of Concordia University in Montréal, Québec, Canada for conducting experiments on glass coated PV modules [2]. The test section with vertical PV modules, was aligned 10° East of South for receiving maximum solar radiation between 8.00 and 14.00 hrs solar time and also avoiding shading from adjacent buildings. The schematic of an outdoor room used for conducting outdoor experiments is illustrated in Figure 10. The outdoor experiments were conducted on a solar wall section built with two commercially available glass coated PV modules, an air passage with gap width of 90 mm, an insulating panel, side walls made up of Plexiglas and connected wooden frames. The insulation panel was assembled with 7 mm thick plywood board enclosure filled with 26 mm polystyrene. The steady state average resistance of the composite plywood board-polystyrene section of insulation panel obtained from the heat flow meter test was 1.0 m^2 °C W^{-1} [2]. The side wall of wooden frame was fixed with Plexiglas of thickness 12 mm (k = 0.1316 W m^{-1} K^{-1}) placed perpendicularly facing air-gap for view of the air passage from inside the test room.

A series electrical circuit connection was established for a pair of vertical PV modules installed on a solar wall for determining the current-voltage measurements and electric power output. The electrical circuit for two PV modules connected in series for generation of electric power with a

rheostat of maximum varying resistance up to 50 Ω is illustrated in Figure 11. A circular variable resistor of 50 Ω was a wire-wound coil with a sliding contact that was used to vary electrical resistance without interrupting the current. The Joule heating at a rate of 73 cal/s (305 W) was predictable at a typical current rating of 2.8 A for a PV module. The null resistance of rheostat was calculated to be 39 Ω at a current of 2.8 A, which was not having any potential difference across it. In Figure 6, the sliding contact of rheostat has a resistance of R_s, with a current ($i= V_r/R_s$) flowing across the circuit; V_r is potential difference across the resistance, R_s.

The passive air ventilation was created in the PV module test section by natural wind or through buoyancy effect in the absence of wind. The active fan pressure was used to achieve higher air velocities by operation of the exhaust fan fixed on opposite façade with respect to the PV test section. The slight negative pressure was induced for drawing low air velocities in absence of wind-induced pressure from the inlet damper into the PV test section through the outdoor room. The exhaust ventilation fan was rated at 270 CFM (ft3/min) (0.127 m3 s-1) at standard atmospheric pressure so as to create between 10-25 Pa of negative pressure depending on the air leakage from the outdoor room.

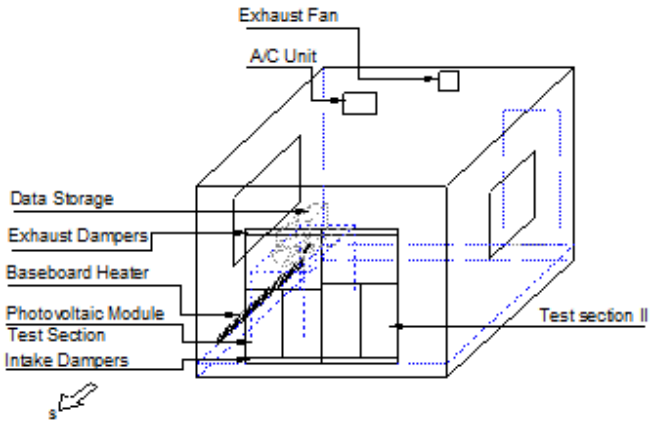


Figure 10: Schematic of a pre-fabricated outdoor room with PV solar wall test section

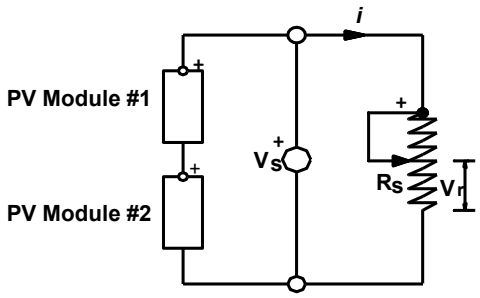


Figure 11: Electric circuit for PV modules

The air velocities near the exit of PV modules were obtained both as a measure of buoyancy and fan induced hybrid ventilation. The fan pressure was created with an exhaust fan fixed on wall of an outdoor room opposite to wall with a wooden frame installed with glass coated PV modules. The air velocity sensor was placed perpendicular to the walls of the PV solar wall section for measuring axial air velocities near its outlet.

In order to minimize boundary layer effect in air velocity measurement, air velocity sensor was placed at middle air-gap width at a distance of 45 mm from either of the walls of the PV solar wall. To minimize turbulence effects in air velocity measurement, air velocity sensor was placed near the outlet (at 1100 mm from the bottom) of the PV test section with total air passage length of 2100 mm in the test section. To minimize three-dimensional effect in air velocities, fan-induced active ventilation was achieved by developing air velocities by the induced draft fan fixed on the opposite wall so that it was placed at a distance from the walls of PV module test section. It was validated by one of the experiments that measured surface temperatures of the two adjacent glass coated PV modules were within ±0.5 °C under similar operating conditions.

7.1. Outdoor Measurements

Three thermocouple sensors were placed at the top, middle, and bottom locations in the PV module, air-passage, and insulating panel. They were used to measure their local temperatures at their defined fixed positions. Two thermocouples were used to measure the inside test room air temperature and ambient air temperature. One differential thermocouple was used to check the air temperature difference between top and bottom sections of the PV module test section. The thermocouple outputs, current, voltage, solar irradiation, and air velocity signals were connected to a data logger and a computer for data storage. The copper-constantan junction (T-type) thermocouples were used for measurements for generating voltage signals to the data logger. As mentioned earlier, the air velocity sensor (make-Dwyer instruments, model no. 640-0) was used to collect air velocity measurements. The solar intensity measurements were collected with a photometer sensor fixed on the south wall of an outdoor room. The schematic of outdoor experimental apparatus for a PV solar wall device is illustrated in Figure 12.

7.2. Temperature Plots

The measurements collected from the sensors were recorded in a data logger as a function of air velocities through a PV solar wall for the cases of passive ventilation and active ventilation with use of fan pressure. Similarly, temperature measurements were obtained from the PV solar wall section as a function of air velocities. The temperatures for PV module, air passage, and insulating panel in the PV solar wall were obtained. The sample measurements obtained from outdoor experimental setup are presented in Table 3.

The temperature plots for PV module, insulating panel, and air for different conditions of ventilation are illustrated in Figs. 13 and 14. The temperature plots with respect to height of a PV solar wall for the data provided in Tables 3 to 5 are obtained. The variation of mean temperatures of PV module, insulating panel, and air in a PV solar wall with approximately steady solar noon irradiation with varying air velocities for the case of active ventilation and passive ventilation are illustrated in Figs. 15 to

16. Temperature differences for various cases of ventilation are provided in Figure 17(a) and 17 (b).

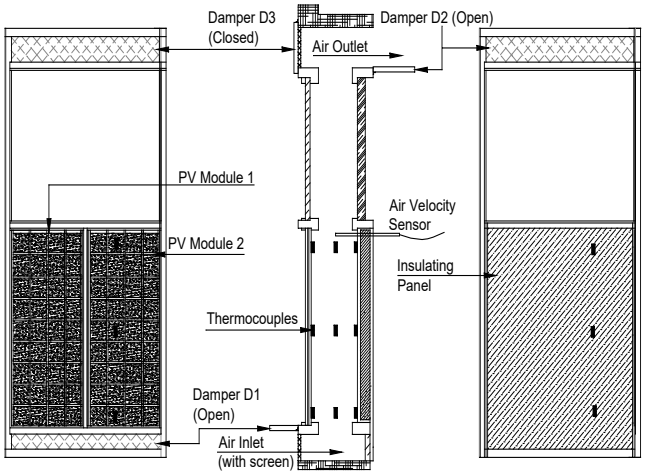


Figure 12: Schematic of outdoor experimental apparatus for a PV solar wall installed on a wooden frame

7.3. Sensible Heat Storage Capacity

The PV solar wall test section with wooden frame was composed of non-homogeneous materials having different densities, specific heats, and thicknesses. The pair of PV modules was fixed with three layers of material, viz. a flat sheet of solar cells, glass sheets on its exterior and interior sides. It was assumed that surface temperature of PV module was uniformly distributed in the three layers. It was also assumed that the heat capacity of the wooden frame and sealing material was having little effect on the temperature of PV module, air or insulation panel. This was because of the fact that wood was used as a construction material, and moreover, the magnitude of the heat capacity of wood framing material was not proportional to the face area of glass coated PV modules. The sensible heat capacities of glass faces, solar cells, air and polystyrene filled plywood board are presented in Table 4. It was observed that the difference of temperatures recorded by the top and bottom sensors for PV module, air and insulating panel were 6.9 °C, 8.1 °C, and 9.9 °C, respectively, for the critical case of passive ventilation of Run no. 4 in Table 3. Furthermore, the temperature differences were used for obtaining sensible heat storage capacities of various components in y-ordinate. The heat storage capacities obtained were 59.6 kJ, 0.755 kJ, and 510.7 kJ for PV module, air and insulation panel respectively. The heat capacities obtained were negligible in comparison with the total daily solar irradiation on PV modules on the day of conducting the outdoor experiments. It was also assumed that there were constant surface properties and ideal still air at the instance of collection of the measurements. Heat storage capacity in x-ordinate was also obtained from the similar procedure by assuming same proportionate temperature difference along thicknesses in x-ordinate. It was found that heat capacity in x-ordinate was nil in comparison with the value of heat capacity obtained for y-ordinate.

Table 3: Outdoor Measurement

Run No.	S ($W\ m^{-2}$)	E_p (W)	T_o (°C)	T_s (°C)	V ($m\ s^{-1}$)
Active ventilation (Fan-assisted)					
1	716.1	30.7	15.2	22.4	0.68
2	716.1	30.7	13.4	22.4	0.53
Passive ventilation (Buoyancy-induced)					
3	697.5	28.9	13.2	25.1	0.13
4	697.5	28.8	13.3	24.9	0.17

Run No.	T_p(b) (°C)	T_p(m) (°C)	T_p(t) (°C)	T_b(b) (° C)	T_b(m) (°C)	T_b(t) (°C)	T_a(b) (°C)	T_a(m) (°C)	T_a(t) (°C)
Active ventilation (Fan-assisted)									
1	35.4	33.8	36.8	20.6	24.7	29.1	18.8	21.7	19.4
2	35.9	34.6	37.9	20.9	25.0	29.5	19.3	22.5	19.9
Passive ventilation (Buoyancy-induced)									
3	40.8	44.9	46.8	27.9	34.8	38.0	21.3	29.5	29.8
4	39.9	45.0	46.8	28.4	35.0	38.3	21.7	28.3	29.8

Distance as per locations shown in Figure 7	T_p(b) (°C)	T_p(m) (°C)	T_p(t) (°C)	T_b(b) (°C)	T_b(m) (°C)
y (cm)	15	55	94	15	55
z (cm)	60	60	60	60	60
x (mm)	6.2	6.2	6.2	96.2	96.2

Distance as per locations shown in Figure 7	T_b(t) (°C)	T_a(b) (°C)	T_a(m) (°C)	T_a(t) (°C)	Air velocity sensor
y (cm)	94	15	55	94	99
z (cm)	60	60	60	60	60
x (mm)	96.2	51.2	51.2	51.2	51.2

Note: x is horizontal; y is vertical; z is adjacent 3rd axis of x-y plane

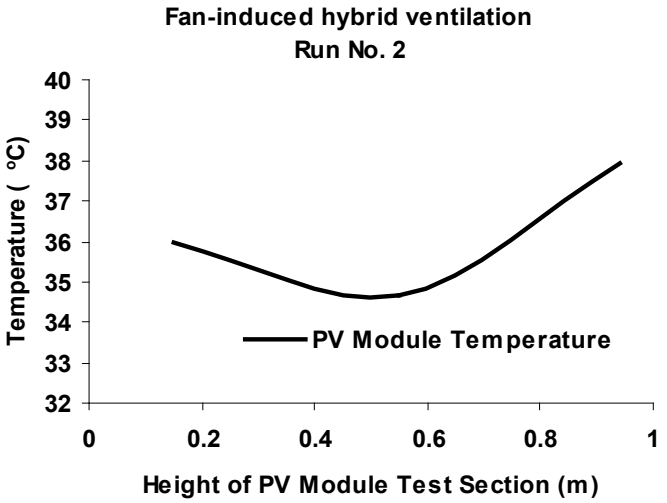


Figure 13: (a) Temperature plot of PV module for fan-assisted ventilation with height of PV solar wall section

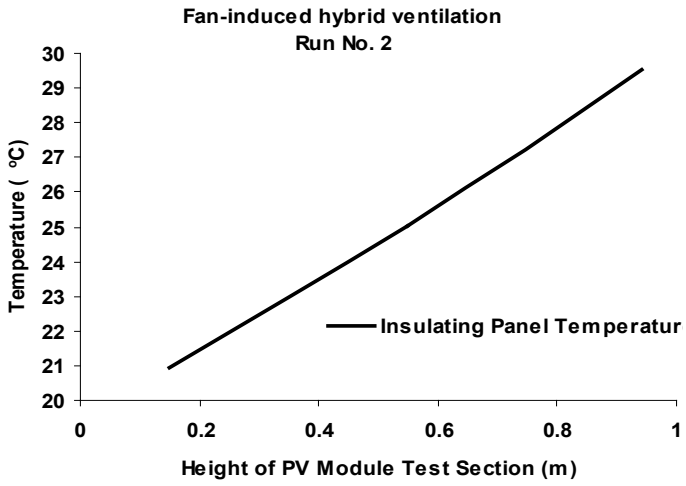


Figure 13: (b) Temperature plot of insulating panel for fan-assisted ventilation with height of PV solar wall section

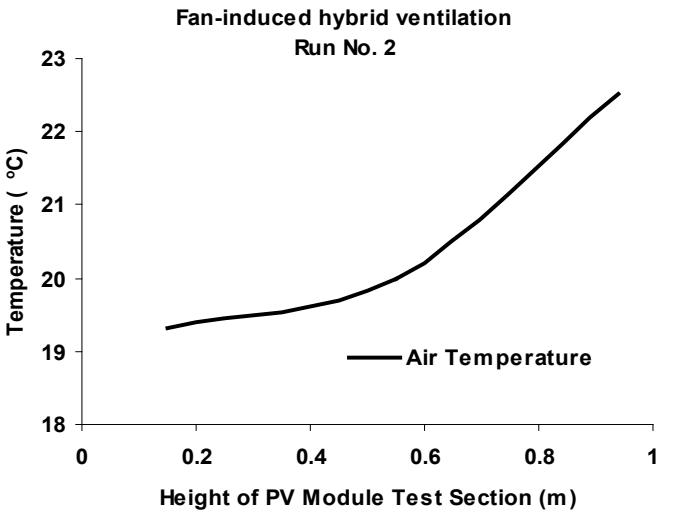
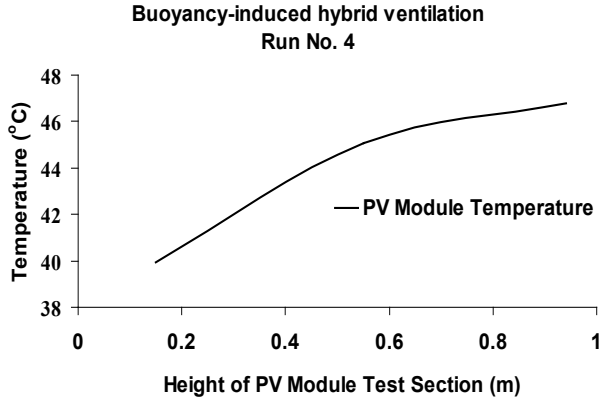


Figure 13: (c) Temperature plot of air for fan-induced hybrid ventilation with height of PV solar wall section

Table 4: Sensible heat storage capacities

Component	ρ_n (kg m^{-3})	C_n (J kg^{-1} K^{-1})	d_n (m) X 10^{-3}	$d_n\rho_nC_n$ (J m^{-2}K^{-1})	H_{pv-T} (J K^{-1})
Glass coating	3000	500	3	4500	4171.5
PV module	2330	677	0.2	315.48	292.45
Glass coating	3000	500	3	4500	4171.5
Sub-total	-	-	-	-	8635.5
Air	1.1174	1000	90	100.56	93.22
Plywood	550	1750	7	6737.5	6245.66
Polystyrene	1050	1200	26	32760	30368.5
Plywood	550	1750	7	6737.5	6245.66
Sub-total	-	-	-	-	42953.0
Total	-	-	-	-	51588.5

Note: Heat capacities were calculated for face area of PV module test section of 0.927 m^2.



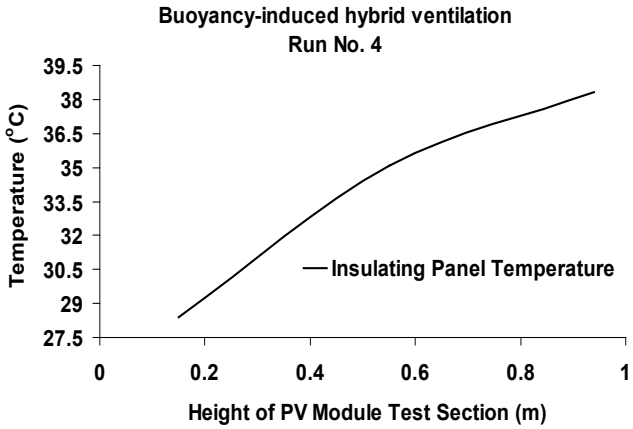


Figure 14: (b) Temperature plot of insulating panel for buoyancy-induced hybrid ventilation with height of PV solar wall section

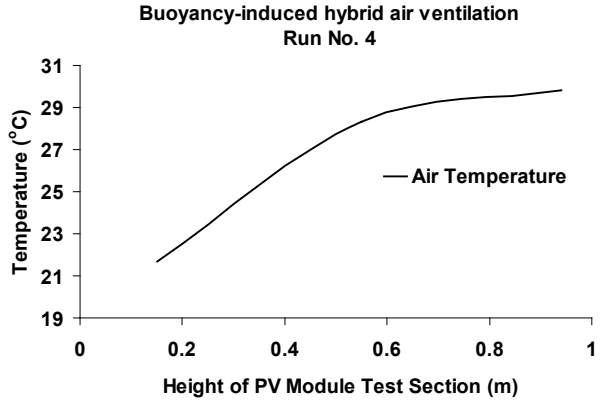


Figure 14: (c) Temperature plot of air for buoyancy-induced hybrid ventilation with height of PV solar wall section

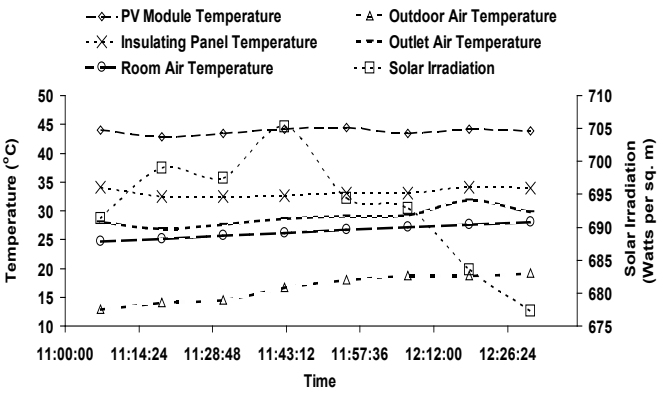


Figure 16: Variation of mean temperatures of PV module, air and insulating panel with solar irradiation under buoyancy-induced hybrid ventilation

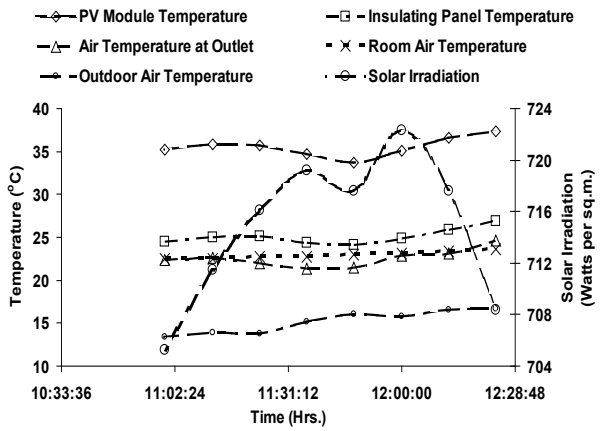


Figure 15: Variation of mean temperatures of PV module, air and insulating panel with solar irradiation under fan-induced hybrid ventilation

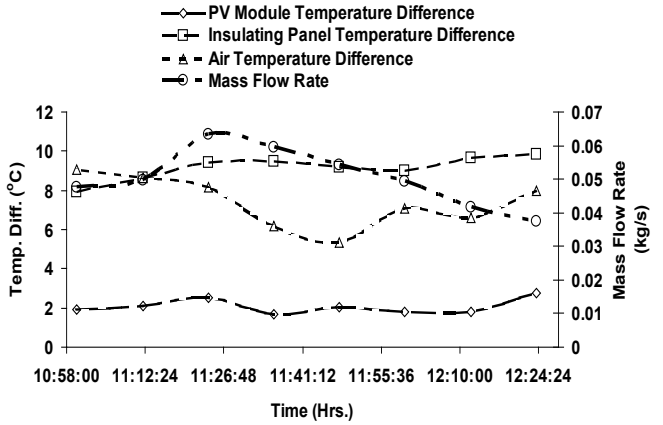


Figure 17(a): Temperature difference for PV module, insulating panel and air with height of PV solar wall under fan-induced hybrid ventilation

Table 5: Thermal storage capacities

Component	k_d (W m^{-1} K^{-1})	$d_n\rho_nC_n$ (J $m^{-2}K^{-1}$)	H_d ($Wm^{-2}K^{-1}$)	T (sec)	ΔT_V (K)	ΔT_H (K)	Q_V (kJ)	Q_H (J)
PV module	0.91	9315.48	10	932	6.9	0.04	5.8	0.2
Air	0.02624	100.56	10.0	10	8.1	0.75	0.0	0.0
Plywood	0.0835	6737.5	10.0	674	9.9	0.40	0.55	0.16
Polystyrene	0.02821	32760	1.0	32760	9.9	0.40	9.0	9.6
Plywood	0.0835	6737.5	10.0	674	9.9	0.40	0.55	0.16
Total	-	-	-	-	-	-	15.9	10.12

Notes:

1. Equivalent thermal conductivity of glass coated PV module was calculated to be 0.91 W m^{-1} K^{-1}
2. Temperature differences along y-direction i.e. along height of PV module test section (0.993 m) were obtained from Table III for Run No. 4 in the case of buoyancy-induced hybrid ventilation.
3. Temperature differences along x-direction i.e. along thicknesses of each component of PV module test section were obtained proportionate to temperature differences along y-direction.

Table 6: Temperature difference and noise of sol with solar irradiation (air velocity: 0.75 $m{\cdot}s^{-1}$)

Solar irradiation ($W{\cdot}m^{-2}$)	Air Temperature Difference (ΔT) °C	Noise of Sol oS (oncisol)
450	15.50	28
550	18.90	28.93
650	22.40	29.7
750	25.90	30.36
850	29.40	30.91

Table 7: Temperature difference and noise of scattering with air velocity (S = 650 W·m^{-2})

Air velocity (m·s^{-1})	Fluid Power (W·m^{-2})	Air Temperature Difference (ΔT) °C	Noise of Scattering oS (oncisip)
1.35	47.62	15.28	17.72
1.05	37.0	18.22	16.50
0.75	26.45	22.40	15.02
0.45	15.87	28.15	12.65
0.15	05.29	29.80	07.64

Table 8: Mass flow rate and noise of therm with (ΔT) °C

(ΔT) °C	Mass flow rate (Kg·s^{-1})	Thermal Power (W·m^{-2})	Noise of Therm oS (oncisol)	(ΔT) °C	Mass flow rate (Kg·s^{-1})	Thermal Power (W·m^{-2})	Noise of therm oS (oncisol)
15.50	0.01376	71.09	19.5602	15.28	0.0231	117.65	21.868
18.90	0.01275	80.325	20.119	18.22	0.0171	103.85	21.296
22.40	0.0120	89.6	20.614	22.40	0.0120	89.6	20.614
25.90	0.0115	99.2833	21.043	28.15	8.1 X 10^{-3}	76.0	19.866
29.40	0.0111	108.78	21.505	29.80	6.2 X 10^{-3}	61.59	18.898

Table 9: Noise of elasticity with air particle velocity (Impedance Z_0 = 413 N·s·m^{-3} at 20°C)

Air velocity (m·s^{-1})	Fluid Power (W·m^{-2})	Noise of Scattering oS (oncisip)	Sound Pressure (N·m^{-2})	Sound Power Intensity (W·m^{-2})	Noise of Elasticity oB (oncibel)
1.35	47.62	17.72	557.5	752.7	30.36
1.05	37.0	16.50	433.65	455.33	28.05
0.75	26.45	15.02	309.75	232.31	24.97
0.45	15.87	12.65	185.85	83.63	20.24
0.15	05.29	07.64	61.94	09.29	10.12

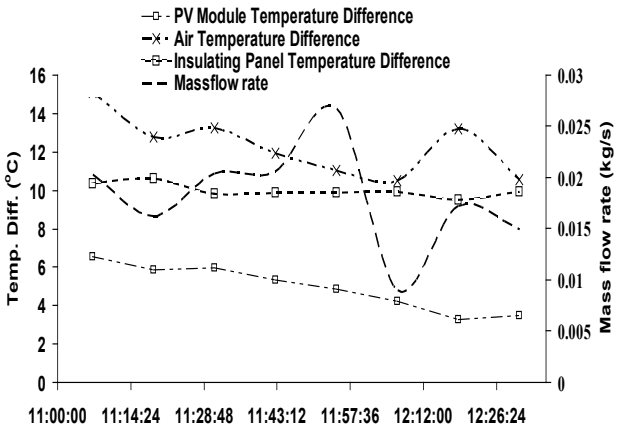


Figure 17(b): Temperature difference for PV module, insulating panel and air with height of PV solar wall under buoyancy-induced hybrid ventilation

8. Numerical Model

The initial boundary values in the model are based on the measured values of quasi steady-state solar irradiation, ambient air temperature, and room air temperature and air velocities. The measured boundary values were then used to obtain solutions for heat equations of the photovoltaic solar wall section. The actual observed temperatures from the experimental apparatus were then compared with the simulated temperatures obtained from the model.

The results of noise unit calculations using Equations 2, 4 and 6 for an outdoor duct exposed to solar radiation are tabulated in Tables 6 to 9. The results of the 2-D numerical model are presented in Table 10 and are obtained by assuming constant fluid and surface properties.

The variation of temperature of air with height of the photovoltaic solar wall test section is illustrated in Figure 18. Figure 18 also illustrates the deviation caused in the proposed model solution in comparison with the ideal exponential or logarithmic temperature variation.

The profiles of surface temperatures of PV module and back panel along the height of the solar wall section are illustrated in Fig, 19. The deviation from the ideal exponential and logarithmic temperature profiles of PV module and back panel is also presented in Figure 19. It was observed that significant cooling of surface of PV module was achieved by radiation heat exchange with the back panel. The surface temperatures of PV module and back panel for four different runs representing variation in solar intensities are presented in Table 10.

8.1. Thermal Time Constant

The thermal time constant is defined as the time required for the outlet air temperature to attain 63.2% of the total difference in value attained in air temperature following a step change in temperature of outdoor air crossing the inlet opening of the PV solar wall. Data for a step change were selected for observing the ambient air temperature. The data selected were in a steady state before and after the time-interval during the unsteady state response of the outlet air temperature with the step change in ambient air temperature. From the graphs, it was observed that thermal time constant under passive air ventilation was between 8-10 minutes in comparison to 2 minutes under active air ventilation. Therefore, it was decided that duration of time

Table 10: Results obtained from 2-d numerical model

Input values	Run no. 1			Run no. 3			Run no. 5			Run no. 7		
S (W m^{-2})	65.1			362.7			520.8			725.4		
E_p (W)	1.2			29.5			31.3			31.1		
T_o (° C)	10.6			11.1			10.8			14.5		
T_s (° C)	15.8			17.9			18.8			22.6		
v (m s^{-1})	0.794			0.597			0.642			0.425		
y (m)	T_a (°C)	T_p (°C)	T_b (°C)	T_a (°C)	T_p (°C)	T_b (°C)	T_a (°C)	T_p (°C)	T_b (°C)	T_a (°C)	T_p (°C)	T_b (°C)
0.0495	10.63	12.97	11.27	11.28	23.66	16.13	11.03	25.67	15.02	15.25	33.87	19.91
0.1485	10.66	12.95	11.57	11.47	23.53	17.19	11.36	25.56	16.14	16.01	33.90	21.57
0.2475	10.69	12.95	11.64	11.66	23.53	17.52	11.69	25.61	16.62	16.76	34.14	22.49
0.3465	10.72	12.96	11.68	11.84	23.57	17.74	12.02	25.72	16.98	17.49	34.45	23.17
0.4455	10.75	12.96	11.70	12.02	23.62	17.89	12.35	25.86	17.28	18.21	34.80	23.80
0.5445	10.79	12.98	11.73	12.21	23.69	18.01	12.67	26.01	17.53	18.91	35.16	24.37
0.6435	10.82	12.99	11.75	12.39	23.76	18.09	12.98	26.18	17.74	19.60	35.53	24.89
0.7425	10.85	13.01	11.75	12.56	23.85	18.12	13.29	26.36	17.89	20.27	35.87	25.32
0.8415	10.88	13.04	11.74	12.74	23.99	18.04	13.60	26.62	17.93	20.92	36.39	25.62
0.9405	10.90	13.09	11.67	12.91	24.25	17.65	13.89	27.02	17.66	21.55	37.01	25.56

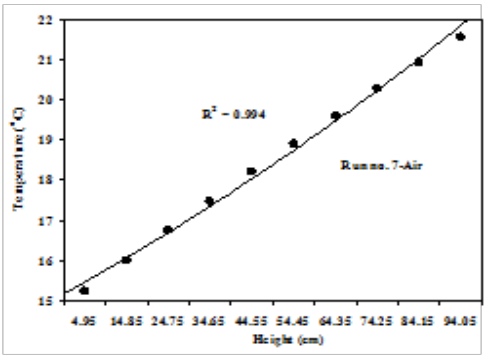


Figure 18: Temperature profile for Air for S=725.4 W m^{-2} (Run no. 7)

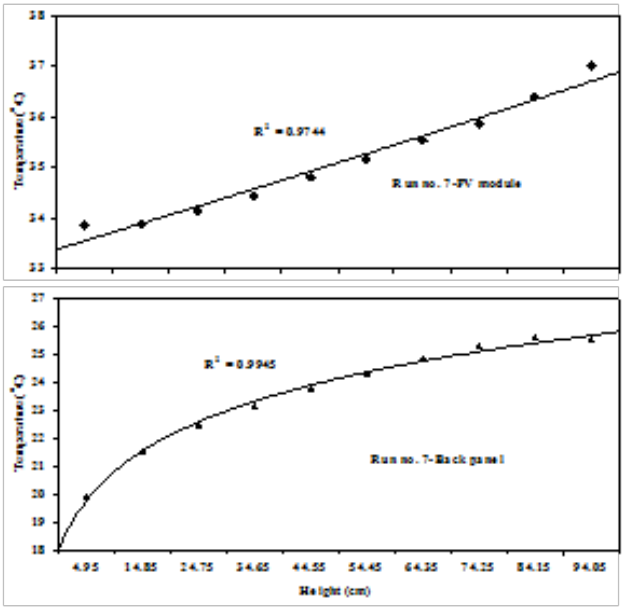


Figure 19: Temperature profile for PV module and Back Panel for S=725.4 Wm^{-2} (Run no. 7)

interval was selected for a minimum of two minutes to record any subtle temperature changes for obtaining measurements from the data logger [34]. The graphs of ambient and outlet air temperatures were plotted against the time-interval of measured data for the two cases of passive and active ventilation are presented in Figure 20(a) and 20 (b). Thermal time constants of the PV solar wall section were function of ambient air temperatures and air velocities and were therefore approximately calculated under conditions of passive and active air ventilation.

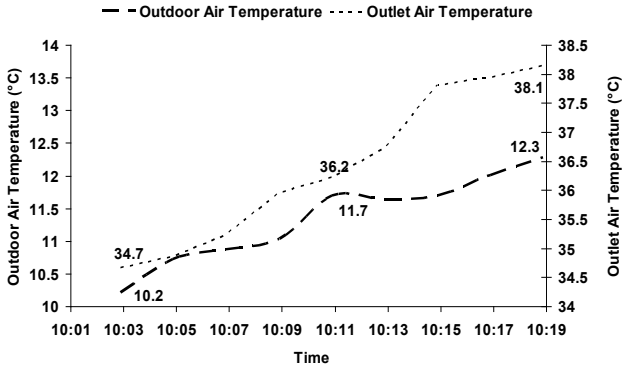


Figure 20 (a): Changes in outlet air temperature from a PV solar wall with a step change in outdoor air temperature under passive (buoyancy-induced) ventilation

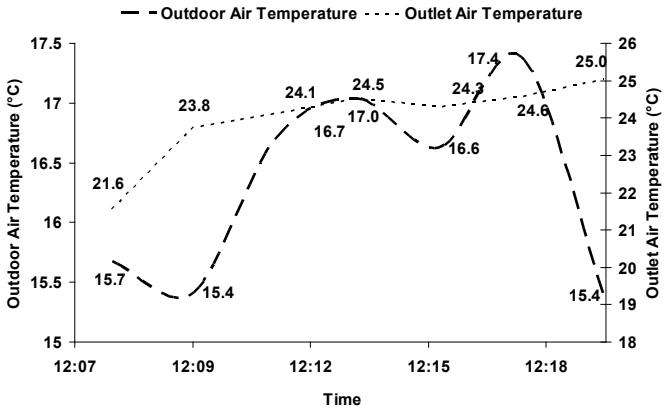


Figure 20 (b): Changes in outlet air temperature from a PV solar wall with a step change in outdoor air temperature under active (fan-induced) ventilation

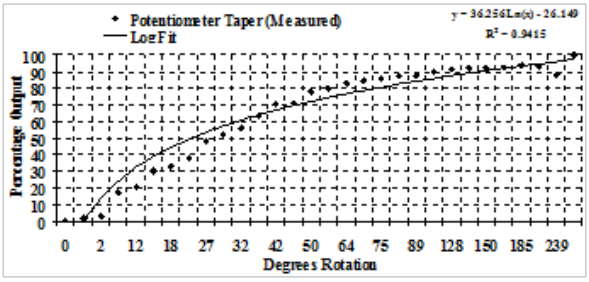


Figure 21: Potentiometer Taper (Measured) with percentage voltage output

8.2. *Photovoltaic Amplification*

In a parallel plate PV device, the temperatures were measured as a function of its volume. The measurements of temperatures for PV modules, insulating panel and ventilated air column in the wooden frame were non-linear. In the ventilated PV solar wall device for transportation of heat, the air velocities were developed both as a measure of buoyancy-induced and fan-assisted ventilation. The measurements of PV solar wall device, ambient air and room air temperatures, air velocities and solar intensities are presented in Table 10. The location and nomenclature of sensors is presented in Table 3. The power output results from a potentiometer with rotation of its circular knob are illustrated in Figure 21. From the graphs of Figs. 13, 14, 18, 19 and 20, phenomenon of photovoltaic amplification is observed. For a PV solar wall device, the gain is a function of its resistance or volume of steady state thermal and electrical functions. A loudspeaker has similar operational characteristics.

9. Signal Processing Parameters for a PV Device

The sinusoidal steady-state response was applied in performing the analysis of the parallel plate PV device circuit, because of the advantage of representing a periodic function in terms of a sinusoidal exponential function. Electrical analog RC circuit parameters of a parallel plate PV device are enumerated as under [4, 22]:

Capacitance: The capacitance of a parallel plate PV device with air as a dielectric medium was calculated to be 91.2 picofarads.

Resistance: The electrical resistances of various components were calculated as: glass coated PV modules were approximated as 5.3 kΩ, air was approximated as 1200 MΩ, and plywood board was approximated as 26.5 Tera Ω. The total equivalent electrical resistance of a parallel plate PV solar wall device was approximated as: 5.3 kΩ.

Time Constant: The time constant, which is product of resistance and capacitance, was calculated to be: 0.5 microseconds. The frequency with this time constant was calculated to be 2 MHz.

Capacitive Reactance: The capacitive reactance was calculated to be: 872.5 Ω.

Impedance: The impedance of the circuit was calculated to be: 5.4 kΩ.

The Phase angle θ: The phase angle between capacitance and reactance was calculated to be 9°.

The Phasor representation:

$Z = 5.300 - j\,0.8725 = 5.4\ \text{k}\Omega \angle -9°$.

Capacitive Heating: The joule law gives instantaneous power absorbed by the capacitive impedance which is converted to heat. The heat capacities under critical operation of buoyancy-induced ventilation were calculated to be 59.6 kJ, 0.755 kJ and 510.7 kJ for PV module, air and plywood board respectively. The total average value of joule heating for the parallel plate PV device was calculated to be 571 kJ.

Induction Losses: The induction losses due to thermal storage effect in the parallel plate PV device were calculated to be 15.9 KJ.

Power Factor: The power factor was calculated to be cos θ = 0.911 lag.

Current function ($i^2(t)$): Using the current function, $i^2(t) = I_m^2 \sin^2(\omega t + \theta)$, the effective (root mean square) value of current was calculated to be 10.4 amps and maximum value of current was calculated to be 14.71 amps.

Voltage function: The voltage function is defined as per the sine wave: $v = V_m \sin(\omega t)$. The effective value of the voltage was calculated to be 60.4 volts and maximum value of the voltage was calculated to be 85.42 volts.

Power function: The instantaneous power is given by the expression:

$$p(t) = \frac{V_m . I_m}{2} . \cos(\theta) - \frac{V_m . I_m}{2} . \cos(2\omega - \theta) \qquad (7)$$

The Plots: The time diagram for current, voltage is plotted in Figure 22 (a). The time diagram for power is plotted in Figure 22 (b).

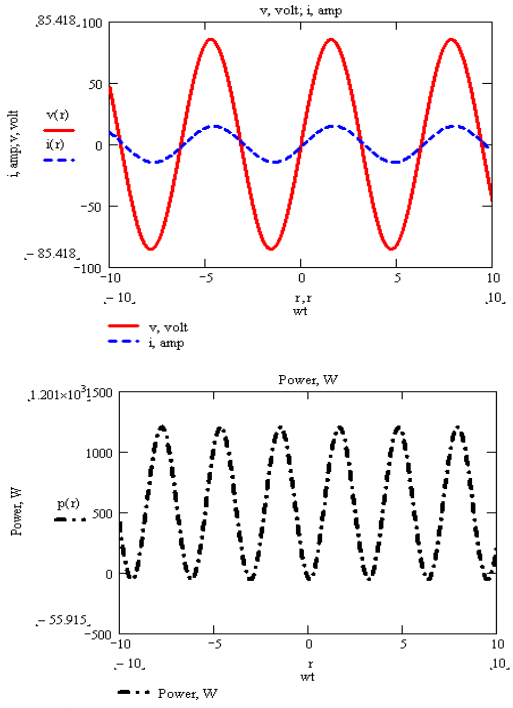


Figure 22 – Time diagrams: (a) voltage and current; (b) power in a RC circuit amplifier

Power Transfer: To describe the resonance phenomenon for

equivalent mechanical, hydraulic and thermal systems of PV solar wall device, an electrical analog was used.

The Figure 22 shows that the instantaneous power is negative whenever the voltage and current are of opposite sign. However, as is illustrated in the Figure 22 that positive area of p(t) energy exceeds the negative area. Therefore, the average power is finite. Since the angle, θ, is small between current and voltage, the negative area of energy become very small. The induction losses due to thermal storage amount to 1.5% in comparison to the capacitive heating. Thus induction losses cannot be avoided in any electrical circuit, but can be minimized.

10. Solar Energy Acoustics

10.1. Theory

A theory of acoustics in solar energy was proposed by the author [10], which can be arrived and deduced from definitions of Noise of Sol and Equations (1) and (2).

A brief on the theory of a solar energy absorber is presented here in this paper. On the transparent surface of a solar energy absorber, direct rays from the sun are incident. The diffuse rays from the sky, clouds, and surrounding objects are also incident, which is also partly reflected and partly transmitted from a solar energy absorber. In addition to this, the selective coating on the surface of a solar energy absorber absorbs some part of the radiation. The reflected part of the incident flux is called the reflectance p, the absorbed part of radiation is called the absorptance α, and the transmitted part of radiation is called the transmittance τ. The sum of absorptance, transmittance and reflectance is unity, or

$$\alpha + \tau + p = 1 \quad (8)$$

There are non-constant distributions of radiations over the directions of incidence and over the wavelength (or frequency) scale on a solar energy absorber. A sample of selective material of a solar energy absorber has radiation properties of transmittance, reflectance and absorptance, which are function of a specific thickness. For a solar energy absorber, the emittance ε is defined as the division of the thermal radiant flux emission from a surface to the black body radiant flux emitted at the same temperature.

The evaluation of radiation properties from angular dependence point of view is explained through geometry theory of a solid angle formed by all rays joining a point to a closed curve. The solid angle is defined as the ratio of the projected area A on the sphere to the square of length R and radius R. A sphere is considered to be having a solid angle of 4π steradians. For a solar energy absorber, the solar radiation incident on a point at a surface is presumed to be coming from many directions in a conical solid angle. For a cone of half angle θ, the solid angle defined by the circular top of the cone and point bottom of that cone is given by expression as:

$$\Omega = 2\,\pi\,(1-\cos\theta) \quad (9)$$

In measuring the transmittance or reflectance, a sample is illuminated over a specified solid angle. The flux is then collected for a given solid angle for measurement of reflectance or transmittance. A conical solid angle is bound through a right circular cone. The source of solar radiation is sunlight.

The radiation properties of sunlight essential for performance analysis of daylighting and lighting are defined as follows:

The luminous flux is directed outward from a source, it ultimately strikes on many surfaces, where it is reflected, transmitted and absorbed. The illuminance is the density of the luminous flux incident on that surface.

The luminance is defined as the intensity of the source or the sink in the direction of an observer divided by the area projected on the source or sink as viewed by an observer.

Luminous intensity is defined as the power due to generation of the luminous flux. A source of sunlight is depicted to be having a luminous intensity in a particular direction. The inverse square law demonstrate that the illuminance incident perpendicular to the line from the point source of sunlight to the surface is directly proportional to the intensity of the source and inversely proportional to the square of the distance from the source of sunlight to the incident surface.

The luminous flux is defined as the rate of flow of light. A receiver surface receives watts of sunlight and it emits luminous flux. The measure of the success rate in converting watts of sunlight to lumens is called efficacy.

Quantity of sources is defined as luminous energy and is related to luminous flux, which is luminous power per unit time.

The luminance exitance is defined as the density of luminous flux leaving a surface of a solar energy absorber. The reflectance is defined as the ratio of the luminous flux reflected from a surface to the luminous flux incident on that surface. The transmittance is defined as the ratio of the luminous flux transmitted through a surface to that incident on the same surface.

10.2. Source and Sink of Solar Energy

The planet earth and its surrounding environment is absorber of solar energy. Examples of solar energy absorbers are civil structures such as green houses, buildings, panels, collectors, thermal power plants, roads, ports, bridges and canals. The earth-atmosphere system is an absorber for a portion of the incident solar energy radiated from the sun. The surface of the earth is also heated by a flux from its interior region from the decay of radioactive isotopes of the earth. The tides on earth occur due to the earth-moon system, result in viscous friction, energy input that affects the surface energy balance of the earth. To the extent that the earth is at a steady state and is not changing, the heat radiated by the surface of the earth is equal to the sum of the heat inputs. The solar radiation sources consist of many waves of interference in its radiation spectrum. The absorbed sources of waves of interference of light, sound, heat, electricity, fluid, fire are function of absorptivity for a solar intensity. The sink of solar energy is the earth and its earth-moon system. The temperature of the earth is raised by absorbed solar radiation as incident solar energy is the dominant energy input. The radiation absorbed by the earth is proportional to the earth's cross-sectional area perpendicular to the sun's flux. The earth is acting as a black body radiator with actual power radiated by the earth is determined

from the radiative temperature of the earth-atmosphere system.

10.3. Radiation Sources and Life on Earth

Life on the earth is completely dependent upon the energy radiated by the sun. By photosynthesis green plants convert solar energy into chemical energy. All biological processes are either directly or indirectly reliant upon photosynthesis. In addition to the primary role of light in a living economy, a constant environment of radiations from various sources produces other effects, reactions and adaptations. These radiations have receptiveness to influence the life activities. The solar radiation, while passing through the atmosphere is reflected, scattered, and absorbed by gas molecules, ozone, water vapour, clouds and dust. The atmospheric length path traveled by sun rays is obtained by the air mass *m*, which is the ratio of the mass of atmosphere to the mass which would exist if the sun was directly overhead at sea level. In this case of overhead earth-sun path, *m* is equal to 1.0.

The sunlight on the earth is the major source of radiations. The spectrum of sunlight constitutes ultraviolet radiation, visible light, infrared rays and radio waves. The x-rays are originated by solar flares and their ionization due to absorption happens high in the earth's atmosphere. From various celestial sources, x-rays also reach the earth's atmosphere. About 60 per cent of the sunlight energy is in the invisible infrared region's indefinite limit in radiation spectrum. Before radiations of shorter wavelengths reach the surface of earth, they are absorbed in the earth's atmosphere. The artificial radioactive elements are produced by bombardment through high energy particles, examples include helium nuclei. The radioactive emanations consist of three components: i) alpha particles, which are positively charged helium nuclei; ii) gamma rays, which are penetrating radiations of very short wavelength but otherwise like x-rays; and iii) beta particles, which are rapidly moving electrons.

10.4. Greenhouse Gases

The most of ultraviolet solar radiation is absorbed by the ozone in the upper atmosphere. The part of the solar radiation in the shortwave region is scattered by the air molecules. The ozone layer is formed high above the atmosphere through absorption of ultraviolet radiation by oxygen. The reversible reaction again turns the ozone into oxygen through absorption of longer ultraviolet rays. The strength of the absorption of solar energy is proportional to the wavelength and absorption bands that are formed at regions of strong absorption. The examples of important atmospheric gases forming part of absorption bands are oxygen (O_2), ozone (O_3), carbon dioxide (CO_2), methane (CH_4), water vapour (H_2O), nitrogen dioxide (NO_2) and chlorofluorocarbons (CFC).

11. Sensors and Transducers for a Human Brain

Your body has feedback systems that regulate the internal environment of your body. The feedback systems make use of storage depots and numerous feedback loops. The monitoring of plasma calcium is a good example of negative feedback. The bones constitute large storage depots for calcium, for the plasma to withdraw these storage supplies in times of need. Our body's homeostatic regulatory systems are represented by feedback loops. The feedback is considered negative, when it is compensating or negates any change. The negative feedback is essential to stabilize a system.

The gastrointestinal tract, the lungs, the kidneys, and skin of your body make exchange of materials and energy between the internal and the external environments. A steady state is achieved by regulatory mechanisms involving the balance between the inflow and outflow of the internal environment that stabilizes the composition of the internal environment. The tendency to regulate the internal environment so that it is maintained in a steady state is called homeostasis [1, 7].

11.1. Heat Stress

When vulnerable to solar radiation, your body acts as a solar energy absorber. This enables your senses for interpretation of our surrounding environment. The amount of heat you lose is dependent on the difference in temperature between the environment and the surface of your body. The heat loss from your body is directly proportional to this difference in temperature. The heat would be released from your body, if the surface temperature of your body is at a higher level than that of the environment. If due to excessive solar radiation, the environmental temperature rises above your body temperature, you will gain heat from the environment.

11.2. Effects of Intense Heat in a Room

Your presence in a room with solar radiation, high room air temperature and conduction do not count in your favour for loss of body heat. In this situation, your body gain heat from the surrounding environment. You have the chance to survive, but only through mechanism of sweating for losing heat. The normal response of your body is resulting in intense heat strains through the body's circulatory system. This happens because the hypothalamus reacts to the heat through expansion of the blood vessels in your skin. As a consequence, it results in decreased resistance to blood flow and your blood pressure falls. The reflexes happen, which prevent large changes in blood pressure, ultimately begin to operate and the decreased resistance to blood flow is compensated through the heart working much harder. The expanded blood vessels create it potential for large amounts of blood to pool in the vessels of your skin at the price of other organs. You will faint as a result of sufficiently low blood supply to your brain.

11.3. Sweating

Through sweating, a circulatory problem occurs, because of the salt and water loss. A decreased plasma volume occurs because of excessive fluid loss. As a result, the output of blood from the heart is slowed down, which could lead to decreased blood flow to the skin, which ultimately could reduce sweating. If this occurred, your main avenue for heat loss would be shut down. In this case, heat production would continue and your body temperature would rise until your whole body system is collapsed. The body's ability to control heat loss is very limited. A vicious circle may occur as

a result of heat not losing rapidly enough to prevent a rise in body temperature. When heat regulation fails, the positive feedback loop occurs into operation in your body (Heat production – metabolism – temperature control); if unchecked it results in heat stroke and death.

11.4. Evaporation

Sweating is an evident method to lose heat by evaporation. Water continuously evaporates from your skin during the process of sweating. When you breathe, there is also a small loss of water from the surface of the lungs. When you breathe or sweat, the amount of water that evaporates is proportional to the humidity of the air. At high humidity of the surrounding air, water evaporates much slowly and therefore contributes less to the cooling process. One of the important methods of losing heat from the surface of your body is through evaporation. There is evaporation of water from your skin after swimming and you feel cool. There should have minimum amount of energy of the water molecules on your body's surface for evaporation to happen. The speeding water molecules on the surface of skin beat the forces holding them in the liquid state. This finally results in evaporation into the air as water vapour molecules. The cooler molecules are left behind in the whole process. As a result, heat pours from the warmer surface of your skin to the cooler water molecules. This heat flow transfers energy to the water, increasing the number of evaporating water molecules. Furthermore, the cooling of your skin also cools any underneath blood which tends to flow through that part of your body.

11.5. Protection through Noise Filters

The keeping of face beard (facial hair) and wearing of a knitted head cloth (*patka*) and a turban (*pag*) on your body has a logical and a scientific significance. The daily self-making folds of hair knots and making round folds of turban over the head of your body with colourful cotton cloths has following historical, medical benefits: i) it indicate, protects and concentrate the disciplinary physical and mental strength of a person; ii) it gives hair tonic to the growth of hairs on your body due to solar energy; iii) the whole system acts as noise filter and provides immunity to your body; and iv) the folded *Patka* with style, folded design of hair knots on top of your head is your identity in time domain, the face beard on your body is a measuring ration and a sign of man, the turban with style, colour, design is your identity in space domain.

12. Comfort and Health

The ASHRAE handbook of fundamentals has provided a detailed discussion of the physiological principles of human thermal comfort [37]. The amount of heat generated and dissipated by your body varies considerably with activity and age as well as with size and gender. The complex regulatory system of your body act to maintain the temperature of your body to about 36.9 °C regardless of the environmental conditions.

The environmental factors that affect thermal balance of your body for influencing thermal comfort are [38]: i) the dry bulb temperature of the surrounding air; ii) the humidity of the surrounding air; iii) the relative velocity of the surrounding air; iv) the temperature of any surfaces that are in direct view to any part of your body and thus exchange radiation. In addition the personal variables that affect the thermal comfort of your body are activity and clothing. The physiological mechanisms that your body uses to control your body temperatures are metabolism, blood circulation, near the surface of the skin termed as cutaneous blood circulation, respiration and sweating. The metabolism rate determines the rate at which energy is converted from chemical to thermal form within your body and blood circulation controls the rate at which the thermal energy is carried forward to the surface of your skin. In respiration, air is taken in at ambient conditions but leaves saturated with moisture and very near the temperature of your body. The sweating has a significant effect on the rate at which energy is carried away from the skin of your body by heat and mass transfer.

The energy generated by metabolism rate of your body varies considerably with the activity of your body. A unit to express the metabolic rate per unit of area of your body is termed as met (1 met = 58.2 W m^{-2}), defined as the metabolic rate for your body while seated quite (called sedentary). The variable which affects the comfort of your body is the type and amount of clothing that you are wearing. The insulation of clothing is defined as a single equivalent uniform layer over your whole body. The insulation value for clothing of your body is expressed in terms of clo units (1 clo=0.155 m^2CW^{-1}). A heavy business suit with accessories has insulation value of 1 clo, whereas a pair of shorts has 0.05 clo. ASHRAE Standard 55 provides comfort conditions for an acceptable thermal environment [38]. The most comfort studies involve use of the ASHRAE thermal sensation scale. The scale relates words describing thermal sensations to a corresponding number. These are enumerated below:

+3	hot
+2	warm
+1	slightly warm
0	neutral
-1	slightly cool
-2	cool
-3	cold

The operative temperature is the average of the mean radiant and ambient air temperatures, weighted by their respective heat transfer coefficients. Energy balance equations are developed that use a predicted mean vote (PMV) index. The PMV index predicts the mean response of a large group of persons. The coordinates of the comfort zones are:

Winter: Operative temperature of 20 to 23.5 °C at 18 °C wet bulb temperature; operative temperature of 20.5 to 24.5 °C at 2 °C dew point.

Summer: Operative temperature of 22.5 to 26 °C at 20 °C wet-bulb temperature; operative temperature of 23.5 to 27 °C at 2 °C dew point.

The loss of hearing is a serious occupational hazard. Factory workers, truck drivers and other workers are exposed to noise levels that are harmful. The continuous exposure to harmful noise

level results in hearing impairment of your ear. The federal Occupational Safety and Health Administration (OSHA) in the Department of Labour has maintained noise exposure limits. The noise exposure is expressed in daily noise for an 8-hour shift.

13. Discussions

13.1. Experiments

The electrical characteristics that were obtained for a pair of glass coated PV modules with indoor experiments under standard operating conditions were performed under ideal conditions. The ideal test conditions gave inconclusive characteristics which certainly are not suitable for consideration during design of PV module system connected to buildings. Additional information under available operating conditions was essential for system design of the PV module system. The results of outdoor experiments revealed that thermal characteristics of PV modules were improved with a glass coating. There was considerable loss in generation of electrical power from a pair of glass coated PV module in comparison to typical rated electrical power for a PV module. The glass coating was responsible for increasing thermal storage along the major dimensions of the PV module system. The conduction heat flow along the height of the glass coated PV module resulted in rise of temperature along its height. This resulted in formation of thermal gradients along the volume of PV module system. The original metallic frame of the glass coated PV module was replaced with a wooden frame to minimize the effect of thermal gradients. The insulating panel in the wooden frame was a plywood board filled with polystyrene which is used in a modern building for insulation. A comparison of experimental results of passive and active ventilation envisaged that active ventilation was necessary to reduce high temperatures of PV modules. The heat storage capacities and thermal storage capacities of various components of PV module system were calculated. The major heat storage in the PV module system was due to heat capacities of the insulating panel. The result was loss of heat along the height of the PV module system. The packed polystyrene with loose fills in the plywood board was responsible for increasing thermal storage capacity of insulating panel. The rise in temperature along the thickness was calculated proportionally from the observed temperature rise along the height of the PV module system. Hence, the temperature measurements were also obtained along the volume, viz. height of a PV solar wall system.

13.2. Numerical Model

The temperature plots of Figure 18 and Figure 19 show that there was steep rise in temperature of insulating panel in comparison to glass coated PV module. The air temperature of air passage in the PV module test section was increased considerably due to increase in thermal mass of the PV module system connected to a building.

A periodic solar heat flux was subjected to heat exchange with ambient and heated air in the PV solar wall. With the assumption of lumped parameter applicability in x-direction imposed a very small percentage of error in the solution of the model. More elaborate solution techniques for heat exchange analysis though improve the accuracy of solution method but certainly were not realistic in environmental conditions involving large deviation due to variation in convection heat exchange coefficients but no deviation with simulated periodic variation of solar intensity.

13.3. Composite Waves due to Stresses and Oscillations

The following composite waves are generated due to development of stresses and oscillations on a PV solar wall device with incident short wavelength electro-magnetic waves: (i) due to connected external electrical load, transmission of electrical energy wave; (ii) due to exchange of viscous dissipation with air, the propagation of heat waves at longer wavelength; (iii) due to thermal stress generation with propagation of heat waves, elastic waves are transmitted in a PV solar wall device; (iv) due to combination of stress development with heat waves, elastic waves and applied external source of energy, the fluid surface waves are propagated; (v) due to climate-particle oscillations of wind and fan induced pressure, applied external waves are propagated; In absence of wind and fan pressure, thermosyphon based oscillations are propagated due to thermal buoyancy.

Such as in an organ pipe, the sound waves can transmit with combination of applied external source of energy and fluid surface waves. From the surrounding environment due to air-borne sound transmission, the sound waves are propagated. Due to various stresses and oscillations acting on a static particle body, the transmission of composite waves is generated. With action of composite forces acting on a PV solar wall device, the developed stresses are classified as: (i) fundamental; (ii) internal; and (iii) external. Due to presence of electromagnetic and gravitational forces of a solar system, the fundamental stresses are generated. Under the influence of fundamental stresses, internal stresses are generated with characterization of composition properties at chemical, atomic and molecular level. With application of external source of energy such as fan based force used for active air ventilation in PV solar wall device, the external stresses are generated. With stress development due to periodic force of expansion/compression, cooling/heating and night/day, the oscillations are assumed to be generated on PV solar wall device. On a PV solar wall device, climate-particle oscillations due to wind force are also transmitted. Due to superposition of composite waves, the fluctuating forces are generated.

13.4. Resonance

The parallel and series cases of LCR circuit resonance are briefed here. With the aid of presented modelling and experimental data, the cases of resonance are visualized. For elastic waves transmission, the inductance force exists due to mass of the mechanical system. The capacitance force exists due to heat storage capacities of PV solar wall device (PV modules, air and polystyrene filled plywood board). The polystyrene filled plywood board is vulnerable to heat stress of fire as soon as heat waves propagated with frequency matching with its latent heat of vaporization is achieved. Due to thermal and fluid resistance in energy storage elements of PV solar wall device, equivalent

electrical analog resistance is developed. The parallel case of LCR resonance happens with fluid surface waves (RC) and heat waves (RC) in conjunction with inductance (L) due to mass of PV solar wall device and resistance (R) due to temperatures of ambient air and ground surface. The series case of LCR resonance occurs with propagation of elastic waves of a PV solar wall device.

13.5. Human Regulatory Feedback Mechanism

Our body has feedback systems that regulate the internal environment of our body. The feedback systems make use of storage depots and numerous feedback loops. The monitoring of plasma calcium is a good example of negative feedback. The bones constitute large storage depots for calcium, for the plasma to withdraw these storage supplies in times of need. Our body's homeostatic regulatory systems are represented by feedback loops. The feedback is considered negative, when it is compensating or negates any change. The negative feedback is essential to stabilize a system. The gastrointestinal tract, the lungs, the kidneys, and skin of your body make exchange of materials and energy between the internal and the external environments.

A steady state is achieved by regulatory mechanisms involving the balance between the inflow and outflow of the internal environment that stabilizes the composition of the internal environment. The tendency to regulate the internal environment so that it is maintained in a steady state is called homeostasis. The coordination of the activities of the various sensory organs within your body is controlled by the secretion of hormones and by transmission of nerve impulses. The nerve impulses travel along axons.

The sensory nerve axons carry impulses from different parts of your body to the central nervous system (brain and spinal cord). Although parts of your brain involved in various physiological functions have been located, research attempts to localize complex behavior patterns in sensory areas of your brain have failed. The sensory areas are located for the impulses on the cortex. The motor areas are also located where "command" impulses leave the cortex. The cerebellum receives impulses from sensory receptors and interacts with motor cortex, insuring smooth muscular movements. The reticular formation receives sensory impulses from the environment and sends impulses to the cortex.

13.6. Noise Filters and Kinetic Theory of Gases

The behavior of gases over wide range of temperatures is predicted through kinetic theory model by average kinetic energy of translation. The noise filters are defined based on the model of kinetic theory of gases for filtering unwanted frequencies of oscillations from a power system. It is a network with selective transmission for currents from a power system of varying frequency. The noise filter is used to filter unwanted frequencies of oscillations from a power system. The noise protection and security is a crucial operation for obtaining a desired output from a power system.

13.7. Significance of Noise Characterization Theory in Cities

Energy & Entropy: Developed cities are highly reliant on continuous inputs of energy to provide goods, services and environment associated with a high standard of well-being. Energy is typically thought of in purely sensible terms of heat and work consumption of a city, but the process development itself may be shown to be theoretically dependent upon inputs of energy in a city. Primarily, energy is required to achieve an increase in order and knowledge (since thought process of an individual requires a metabolic input). The theoretical basis for this concept of increase in order and knowledge is based on principle of kinetic theory of gases. The laws of mechanics are insufficient to predict the behavior of a gas, since they provide no preferred direction for time. With some limitations, the concept of increasing entropy, however, provides a direction for time. Since the entropy of a system is associated with its lack of order, the concept of entropy also applies to non-thermodynamic systems [39].

Biological System and Entropy: Biological system represents an increase in order for the evolved material (i.e. a decrease in entropy) and is termed as 'complexification' by the biologists. Cultural evolvement and industrialization also signify movement by a developing city from less order to greater order. Additionally, life also requires continuous input of low entropy energy to recompense for the ultimate entropy increase experienced by all living material, i.e. death.

Photosynthesis is directly reliant upon energetic photons of shorter wavelength visible light. By means of photon-activated chemical reactions, plants are able to produce organic material from carbon dioxide and water. Since the energy potential of the organic material is greater than that of constituents, an entropy decrease is achieved. The organic material has higher order degree than the carbon dioxide and water from which it was produced. The process of respiration both by the plants and living organisms which depend either directly or indirectly through plants, achieves in low-temperature waste heat. It is this biological heat engine upon which life on earth is almost entirely dependent.

Knowledge, Order and Well-Being: Knowledge is gained mostly through social-economic process, which requires inputs of low entropy reserves. Well-being of a city also increases through knowledge even though history has been marked with short-term reversals such as in the event of a war. Smart improvements in a city represent an increase in order i.e. a decrease in entropy. Improvement in well-being of a city is often the result of endeavours that have very subtle effects such as an insight of a new scientific theory, which represents a more ordered image of the world. In addition to increases in order, human created systems frequently produce undesired forms of disorder, i.e. environmental pollution. For example, emissions (indoor and outdoor) represent a 'scattering' to the winds of pollutants, an increase in disorder.

Solar Energy and Waste Recycling: Recycling of wastes is equivalent to interconnecting the material inputs and outputs. Energy though conserved cannot be completely recycled, due to inevitable increase of entropy. A shift from the present dependence upon conventional energy resources to either a direct

or indirect usage of solar energy is already happening. A complete recycling of materials and a total reliance upon solar energy would lead to sustainable and smart economy. Energy derived from the sun is obtained without entropy increase. There is generation of waste heat irrespective of energy source. However, a solar energy system results in no greater increase of entropy in comparison which occurs without the system. Entropy is an indicator of non-utilization. The increase in entropy specifies a decrease in low entropy sources of energy with high utilization, which occurs irrespective of energy conservation in a system.

Noise and Noise Behaviour: It is recognized that many proposed methods of energy conservation also results in reduction of energy cost. For example, an improved transportation system would not only result in energy conservation, economical affordability but also generates less Noise. As is the case of increased energy consumption, increased quantities of goods and services result in generation of more noise behaviour in a city. As stated previously, 'Noise' is defined as a sensation of unwanted intensity of a wave. It is perception of a pollutant and a type of environmental stressor. An environmental stressor such as noise may have detrimental effects on various aspects of health. The unwanted intensity of a wave is a propagation of noise due to transmission of waves (viz. physical agents) such as sun, light, sound, heat, electricity, fluid and fire. Noise Behaviour is checked by identifying a source and a sink of noise i.e. a person making noise in the environment and a person affected by such noise in the environment.

The omnipresent characteristic of the urban environment is its exposure to environmental noise. The excessive noise is accepted by the public health doctors and professionals as an undesirable feature of the urban environment. Noise is indisputably perceived to be an irritant, interfering into personal privacy, and causing displeasure and diminishing the worth of a person and his health. As health is largely defined to include quality of life and displeasure, therefore noise undoubtedly affects health.

The effects of stress due to noise of physical agents are consideration to contribute to a range of disorders as wide as impairment of hearing, cancer, heart disease, musculoskeletal conditions, skin disease, gastrointestinal and other disorders. The evidence is strongest for links between certain types of prolonged stress and heart disease, hypertension, and mental illness. Most of the evidence for such links is epidemiological. It is possible that what is bad for the individual employee is also bad for the organisation. Organisational worries connected with work related stress include high absenteeism, increased staff turnover, low job satisfaction, low morale, poor organisational commitment, poor performance and productivity, possible increased accident and near miss rates, and, in some cases, an increase in employee and client complaints and litigation.

All such unwanted disturbances of noise caused by a person or deviation from a normal behaviour so as to distract attention of a normal person are termed as 'Noise behaviour'. There is always a source of noise and a sink of noise i.e. a person making the noise in the environment and a person affected by such noise in the environment. A person making the noise in the environment is also affected. The occurrences of noise related stress alters the way people think, feel, and behave. Many of the changes that occur are diffident and potentially reversible, although harmful to the person's value of life at the time. Other changes may be more enduring, and have substantial consequences for health. Behavioral changes include increases in health risk behavior, such as smoking and drinking, and decreases in health positive behavior, such as exercise and relaxation. Many behavioral changes represent attempts to cope with the emotional experience of stress due to noise—for example, by making noise due to presence of physical agents in the environment or because of other person's noise behavior in the form of polluting the environment. The noise in the environment is associated with poor decision making, impaired concentration, reduced attention span, impaired memory, and confusion. People who report under stress due to noise also admit that not being able to think straight forward. Social behavior and interpersonal relations of the person affected by noise may also be demolished, possibly reflecting these and other psychological changes such as tiredness and increased irritability.

Biomedical Instrumentation for a Noise System: Biomedical instrumentation for measurement and monitoring of noise from human systems in a city require real time informatics capabilities. Sensing and actuating capabilities as well as measurement systems for noise can be 'in vivo' and 'in vitro'. The signals can be classified as bio-electric, bio-sound sample, bio-mechanical, bio-chemical, bio-magnetic, bio-optic and bio-impedance depending upon origin of 'stresses and oscillations' in a noise system.

Energy Conversion, Energy Processes and Noise: Energy conversion materials and devices convert one form of energy into another form. The energy conversion efficiency of a device is an expression of energy balance between energy input and energy output. Materials play important role in conversion efficiency due to involvement at molecular and atomic level. Energy conversion is deeply associated with energy processes of heat, fluid, electricity and HVAC devices.

Combustion is primary form of energy conversion of low grade fuels such as biomass and municipal solid waste. Energy conversion of low grade fuels from solid to a liquid or gaseous state is through various techniques for enhancement of calorific value of fuel. Inefficient fuel combustion is wasteful as well as generates higher level pollutants. The pollutants and solid wastes generated during energy conversion must be controlled to protect public health and well-being of the affected environment in an ecosystem.

Energy conversion efficiency can be increased through cogeneration. A cogeneration plant consists of energy equipment to produce both electric energy and useful forms of thermal energy such as heat or steam for commercial, industrial, heating, or cooling purposes. Cogeneration plant is designed as either topping-cycle or bottoming-cycle process. Either of these cycles can utilize thermal energy to meet process heating, ventilating, and air-conditioning (HVAC) or built environment comfort requirements for steam or for hot/cold water.

There is generation of waste heat, entropy and noise in all energy conversion devices and systems, which can be minimized by selective materials, optimized design configurations and performing high-resolution modelling and simulations.

13.8. Significance of Noise Characterization Theory in Acoustics

Traditionally acoustics is associated with sound. Nearly all life activities result in generation of sound. However, in traditional acoustics only vibrations by way of pressure difference is associated with sound. In noise characterization theory, apart from sound, acoustics is associated with composite wave elements of light, heat, fluid, electricity, fire and sun. Since all life activities are associated with these composite waves, therefore contemporary acoustics should also involve these wave elements. In this way, knowledge of wave theory principles enables the requirements for the control of noise due to these composite wave elements.

A clear differentiation must be made between the objective specification and measurement of noise field due to these composite wave elements and its subjective appraisal or the sensation it produces on people exposed to it. Many kinds of subjective judgment are possible, according to the particular circumstances, but no objective measurement can determine subjective response directly. Although much of the discussion of noise can be in objective terms and be concerned with quantities which can be evaluated and measured, it is essential to introduce the subjective aspect since noise is associated with people and their routine lives.

Investigation of the noise quantities due to composite wave elements vary with frequency. Frequency is normally plotted on a logarithmic scale because the pitch, which is a subjective assessment of frequency of composite waves (viz., physical agents) perceived by human senses, varies approximately in this way. Usually, measurements are normally made in ranges of frequency bands, or at frequency intervals which are equal steps on a logarithmic scale. Octave steps may be suitable in many cases of composite wave elements, but narrower bands may be more appropriate where applicable. For comparison purposes, certain standard frequencies and frequency bands may need to be developed for these composite wave elements.

For some cases, such as expressing the reduction or insulation provided by a particular building wall, the relative values in *oncibel* and *oncisol* are sufficient. A wall with reduction in sound power intensity and thermal power intensity is given by a certain ratio. Sound pressure measurements are normally made in terms of Root Mean Square (r.m.s) value of the alternating pressure of the sound wave. Similar r.m.s values can be obtained for light, heat, fluid, electricity, fire and sun. Objective measurements and specification of intensity level is to be made on a hexagonal noise scale with seven edges. Figure 23 has presented a double-sided hexagonal slide rule with seven edges for noise measurement representing seven sources of noise [6]. In subjective assessment, however, we are concerned with the way people perceive and respond to various physical agents, and the response of human senses play important role. Contours of equal loudness (pressure level) can also be determined for a number of similar conditions using bands of noise and perceiving conditions of physical agents more closely resembling under practical conditions. A number of other criterion curves, similar in form to the equal contours, are in use for assessment of subjective attributes of noise, such as noisiness, disturbance etc. A simple subjective assessment is presented in Table 11, in which grades and flag colors of noise are notated under limiting conditions [6].

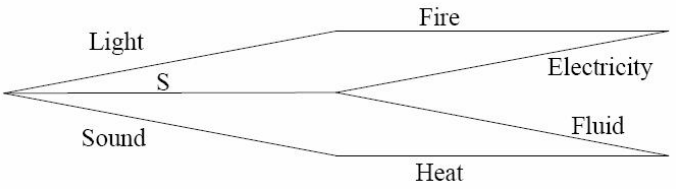


Figure 23: A Double-Sided Hexagonal Scales of Noise with Seven Edges (S denotes Sun)

13.9. Significance of Noise Characterization Theory in Electrical, Instrumentation and Control

Electrical noise is defined as unwanted signals that interfere with the wanted signal output. In electrical engineering, noise is associated with circuits, devices, power systems, electronics, instrumentation and control.

There are many sources of measurement noise. It is mainly caused by the sensor physical mechanism. In general, for sensors and transducers, there are many sources of noise that lead to: i) Johnson's noise; ii) radiation noise; iii) shot noise due to recombination; iv) flicker or contact noise; and v) interference noise.

A control system is gained in measurement noise by way of amplification and it is fed into the whole system. This results in substantial variations in the signal for control of process variable. In this way measurement noise, disturbs the information about the process variable which is delivered by the sensor. Measurement noise has high frequencies in most cases. It is essential that the fluctuations in the control signal are maintained to minimum and not so large to cause the actuator to steep. Measurement noise and its actuator steep leads to inevitable gain of high frequency in the controller and therefore results in its rapid response speed.

In computer controlled systems it is also caused by the resolution of the analog to digital converter. The signal to noise ratio (SNR) is plotted for example with a Bit Error Rate (BER) in a digital communication system. The units expressed in various types of electrical noise should be in *oncisol* which can be plotted on a log scale against frequency or any other parameter. Similarly noise due to an actuator for example a stepper motor, can be expressed in terms of electrical noise (*oncisol*) and its motion noise (*oncisip*).

13.10. Noise Fields

In a free space, noise travels away from a point source, and is small compared with the wavelength. In this case, there is spherical radiation and the intensity at a remote point varies inversely as the distance from the source. There is another case

when noise source is large or extended and the distance is not large compared with the noise source size, in this case the inverse square relation does not hold. Noise fields are complex pattern. Wave theory is necessary to estimate the complete behaviour of noise and even in simple cases, analysis is complex. In actual field practice, only geometrical and statistical methods are used.

Table 11: Noise Grades and Flag Colors under Limiting Conditions

Grades	Noise Grades and Flag Colors under Limiting Conditions		
	Noise of Sol	*Noise of Scattering*	*Noise of Elasticity*
$G_2^a = \pm U$	Sol	Sip	Bel
$G_1 = G_2 = U$	No Positive Solar Energy	No Positive Fluid Energy	No Positive Sound Energy
Base Color for $G_1 = G_2$			
$G_1 = U+\rightarrow 0\ Wm^{-2}$	Decreasing Solar Energy	Decreasing Fluid Energy	Decreasing Sound Energy
Base Color for G_2			
G_1= + ve	Increasing Solar Energy	Increasing Fluid Energy	Increasing Sound Energy
Base Color for G_2			
$G_1 = -U\ Wm^{-2}$	Negative Solar Energy	Negative Fluid Energy	Negative Sound Energy
	Darkness	Low Pressure	Inaudible range
Base Color for G_2			
G_1 = -ve	Darkness increasing, distance from point source of light increasing	Low pressure increasing, vacuum approaching	Inaudible range increasing, vacuum approaching
Base Color for G_2			
$G_1 = -U+\rightarrow 0\ Wm^{-2}$	Negative Solar Energy	Negative Fluid Energy	Negative Sound Energy
	Decreasing Darkness	Decreasing Low Pressure	Decreasing inaudible range
Base Color for G_2			

a. Reference value of $G_2 = \pm U$ signifies the limiting condition with areas of noise interference approaching to zero.

We need to define the system boundary in a noise system. Noise is radiated initially from a source, but due to finite dimensions of the system and as per wave theory some portion of the noise is reflected, whenever it strikes a bounding surface. The resulting noise field at any other point is made up partially of noise radiated directly and partially of the reflections from the system boundaries. When the system surfaces are made as reflecting as possible and the system boundary is irregular with reflections occurring in all directions, the noise field is effectively diffuse. There is no preferred direction in which the noise is travelling at any point and the noise pressure level is assumed to be uniform throughout the volume of the system. There is exception when the noise pressure is not uniform, when the point investigated is close to the noise point source.

Materials play important role in noise absorption. Most of the materials available are not complete reflectors. In many cases, noise is partially reflected, partially absorbed and partially transmitted through the system surfaces forming the system boundaries. The magnitude and direction of noise field depends on the material and surface characteristics of the system boundary. Therefore, the noise fields occurring in practice are rarely completely diffuse. In most of the cases, in the system boundary, there are surface irregularities which are sufficient to provide diffusion at the higher frequencies, where wavelength is short. In the cases, where the noise source wavelength is comparable with the dimensions of the investigated system (such as a cubical or rectangular), complete diffusion may occur at lower end of the frequency range. For many practical cases, it is assumed that the noise fields in the system are diffuse, with support of the measurements made at a number of different positions throughout the investigated system under steady state conditions, which do not show much variation, hence the average of a suitable number is normally taken.

Insulation and Absorption: Physical agents which are propagated into the system fall into two classes, wanted and unwanted. The wanted physical agents are the new field of acoustics. The unwanted physical agents, noise, are controlled in a system in two ways, by insulation and by absorption. Energy is removed from the system containing the sources of physical agents by the absorption at the bounding surfaces. The steady state level to which the physical agent builds up with a given source is determined by the volume and total absorption of the system. Small proportions of the non-reflected composite wave elements of physical agents, which are incident upon a system surfaces, are transmitted through the structure of the material and may be radiated from the other side of the system. In this case the surface acts as an insulating barrier between the source of physical agent and the remote position. Materials which are good absorbers of physical agents are rarely effective insulators if used in the form of barriers in a system. However, the principal requirements for each function are completely different. Thus it is important to distinguish between the two mechanisms by which intensities of composite waves are reduced. The reduction of noise field and kinds of materials suitable for insulation and absorption can also be investigated.

When an insulation panel acts as a barrier to a noise field the practical insulation or reduction to intensities of composite wave elements is given by the amount by which the noise pressure level due to a source is reduced by the barrier. The surface densities, thermal, optical and elastic properties, of the insulation panel are direct factor influencing insulation. It cannot be expected that this simple statement will take into account all the factors which are likely to affect the insulation of a panel in practice. However, it gives an extremely good indication of the values likely to be expected. The noise insulation expression is required, which gives a good indication of the insulation to be expected from an insulation panel at sufficiently low and sufficiently high frequencies depending on the particular physical characteristics of the panel. Between these frequencies there is a band where the material stiffness and the corresponding bending wave pattern in the panel influences its noise insulation behaviour. The panel is

forced by the noise field into bending waves of this wavelength and at the same frequency as is present in the incident composite wave. If the physical constants of the panel-stiffness, density, thickness, optical and thermal characteristics, for example – are such that a free bending wave of this wavelength coincides with this frequency, there is theoretically the possibility of total transmission, with the panel, being freely excited and radiation from the other side occurring. At a particular frequency the bending wavelength imposed on the panel is smallest for noise fields near grazing incidence and largest towards normal incidence. At normal incidence the imposed bending wavelength is infinite and true bending waves cannot be excited. In this case the panel can be forced into uniform oscillations of physical agents. This effect is only possible above the critical frequency for which the bending wavelength in the wall is equal to the wavelength in contact medium since only then is the bending wavelength larger than the wavelength of contact medium. It is simply experienced in practice that this effect has some influence on the insulation values of panel when the appropriate conditions are satisfied. Although noise fields in a closed system (such as rooms) are reverberant and partitions are not normally exposed to plane composite waves, at particular angles of incidence. Although the transmission is never total nor the insulation reduced to zero (because of finite dimensions, the damping in the material and at the edges, and the random noise fields which exist in practice). The depression or dip in composite wave which is obtained can have a significant subjective effect. One way of increasing the critical frequency to above the practical range and hence increasing the insulation is to reduce the stiffness of a panel.

Radiation from Panel: The power of panel radiating noise field is also closely connected with the critical frequency. The depression or dip in the behaviour of insulation occurs because noise is freely radiated when the wavelength in the material is greater than the corresponding wavelength in contact medium. At frequencies for which wavelength in contact medium is greater than wavelength of the material, the value of incidence angle is imaginary and radiation no longer occurs, the oscillations of composite waves in the panel occurs without much stress or pressure field. On the basis of this radiation theory it is expected that thicker panels for which wavelength in material is larger would radiate better than thin ones, and this is so when the amplitudes of oscillations are the same. For the same amount of energy used to excite a panel, however, the thick panel has a lower amplitude of oscillation but better radiating power and thin panel the higher amplitude and the lower radiating power; the net effect of a given excitation, is to have the same order of noise.

Factors affecting Insulation: There are a number of interacting factors which affect the insulation. Mass is normally of greatest importance and determines the insulation at low frequencies. At the lowest frequencies it is possible that the normal resonances or modes of oscillation of panel of typical size may affect the insulation but the practical difficulties of determining the insulation at low frequencies are such that the measurements are not as reliable or reproducible as measurements at higher frequencies. Therefore, any effect caused due to low frequency resonance is difficult to detect.

Heat transmission also plays important role in thermal insulation [40]. The three types of heat transfer affecting the comfort level is radiation, conduction and convection. The integrated thermal and sound insulation is to be determined by various factors such as climate, heat transfer rate, solar energy intensity, elastic, thermal and optical transmission properties. Water vapour barrier and insulation play important role in comfort and wellness of the occupants. In warmer climates, water vapour barriers on exterior side of panel or wall are normally not provided because: i) the temperature drop across the insulation is usually less, and because of this vapour migration would be slower; ii) there are less chances of condensation due to less temperature difference; iii) the built environment is subjected to significant amount of cooling in summer, the vapour barrier would be on the wrong side of the panel much of the yearly time. This is because when the vapour barrier is on the cold side of the insulation, the barrier traps moisture inside the wall. A good integrated insulation will act as noise filter to prevent unwanted oscillations of composite wave elements. Other factors affecting insulation of a system are infiltration. Air and moisture leak into the closed system (or building) through cracks around openings such as windows and doors. This leakage is caused by a combination of the wind pressure on the system and the difference in temperature between the inside and the outside. Sound insulation will depend on the materials and methods used in a system. As mentioned before sound transmission through a panel depends on mass of the panel and on its inelasticity. Massive, thick walls provide excellent sound barriers, but there is a diminishing point when mass alone is not an economical solution.

Integrated noise insulation ratings depend on materials and assemblies of materials. Integrated Noise Transmission Class (INTC) can be used to represent the transmission loss performance of the panel construction over a broad range of frequencies. The results can be plotted, and a standard curve based on human senses response at the various measuring frequencies can be used to determine the rating. The higher the rating, the better will be the construction of a panel in terms of reducing composite wave elements. Average Integrated Transmission Loss (AITL) can be used to check performance of the panel over a range of different frequencies. The results of these performance tests, at each frequency, should be averaged together for integrated transmission loss rating. In these performance tests, the effect of various people sensitivities to various types of noise over various frequency ranges is not taken into consideration. Integrated Noise Reduction Coefficient (INRC) rating could be used to measure the ability of a material to absorb various physical agents rather than reflecting them.

Double or Cavity Walls: When a single panel or wall does not provide sufficient insulation and it is impracticable either to increase the weight or to gain adequate advantage from single wall or panel, a system of double wall construction could be used. The double wall insulation obviously depends on the overall behaviour of each panel, on their relative spacing along with their structural joints for support and interconnection. In the case of

double wall, the air cavity acts as a coupling. If the air in the cavity between the two panels of a double wall is prevented from moving laterally, i.e. along the width and height, rather than across the cavity, the lower frequency resonances due to noise field at oblique incidence and the resulting effect on the insulation are less noticeable. Such a lateral restriction occurs whenever a suitable absorbing material is put into the airspace. The effect of such materials is most noticed at the lower and upper ends of the frequency range and their influences least at mid-frequencies. The best type of material should have a reasonably high resistance, good damping, and stiffness adequately high but not higher than the air. Complete isolation of the two panels of a cavity wall is normally impracticable. Even when the two adjoining panels are capable of standing independently it is normal practice to link them together for structural stability. The way in which these links provide integrated thermal and sound bridges between the panels is likely to have a marked effect on the resultant integrated insulation. Sufficiently light or resilient bridges on the one hand or sufficiently heavy or rigid bridges on the other are likely to have least influence on the insulation, since in either case the impedance of the bridging element, i.e. resistance to being driven into oscillations due to composite wave elements, is different from that of the two panels. It is difficult however, to ensure that a bridge used in practical situation, falls into either of these two categories, hence the use of bridges in cavity wall structures is likely to result in lower values of integrated airborne sound and thermal insulation than expected, and should be avoided if at all possible in practical situations.

Building Structures: In addition to being excited by various airborne physical agents, building structures can also be directly excited by direct impact forces, oscillating bodies such as machine vibrations and footsteps. These forces can also be natural such as wind, thunderstorm, earthquake, lightning etc. One difference between these two methods of excitation is the effective area of the structure which is excited. With airborne physical agents, it is normally the whole area of wall or surface exposed to the noise field, unless the source is particularly directional and these are marked differences in level between the direct and reverberant noise field exciting the structure. With direct impact forces, the area directly excited is more limited, although the structure-borne waves produced may travel through the structure and be radiated from larger areas remote from the noise source. The degree of excitation close to the source may, however, be much greater than with airborne physical agents at the levels normally experienced.

Direct structural excitation may frequently be considered to act at a point and the amplitude of oscillation in the immediate area may well be large although when the energy is spread to other parts of the structure. The resultant amplitude of oscillations is of the same order of magnitude as with airborne composite wave elements. The obvious solution for most impact force transmission problems lies in preventing the impact stresses or pressures, which give rise to oscillations in the main body of the structure, either by providing surfaces which act as cushion absorbers, or by having a discontinuous structure preventing the oscillations implanted on the noise source side from being transmitted to and radiated from surfaces beyond.

With thicker constructions the local elasticity at the point of impact of physical agent may be such that the simple expression for the velocity cannot be used since the noise fields from all types of physical agents not on their velocities but on the resultant velocity subsequently produced in adjacent regions, and on the local resilience at the point of impact. The radiating power of the surface excited also has an effect and this in turn depends on the ratio of the frequency to the critical frequency of the panel. Further, the level of radiated noise field, which can be obtained from the resultant velocity, assumes a constant bandwidth with respect to frequency whereas practical spectra are normally obtained in terms of constant percentage bandwidth.

Design and Insulation: Ventilation systems such as an airflow triple glazed window with an automated roller blind and photovoltaic solar wall which include adequate noise attenuation can be provided in new building construction. At present they are expensive, and the problem of adapting existing buildings for such type of acoustic filter is difficult though not impossible. In particular situations where noise due to physical agents is intermittent, automation of the integrated noise and ventilation system could be performed. The provision of adequate noise insulation in a new building is always possible providing the standards which have to be achieved are known and the appropriate decisions are made at the planning and design stage. In many cases the appropriate standards can be met without extra expense at planning and design stage but it is extremely difficult to make any appreciable changes when the building is already in existence, particularly when high values of insulation are required. At the higher level of insulation, however, the effect of flanking transmission via the supporting structure is most marked and slight changes in structure may give significantly different results. It is at these higher levels of insulation that the results of field and laboratory measurements of noise insulation show the greatest difference. If a completely new form of construction is planned for which no corresponding field results are available, a laboratory measurement will indicate the potential insulation which the construction could give.

Maximum intruding noise levels: In every case, before the acoustic design is considered, the problem of eliminating external noise has to be solved, removing the unwanted physical agents, so that wanted composite wave elements can be perceived to the best advantage. The principles of noise insulation has been described, but in practice an appreciable proportion of the overall design is concerned with ensuring that these principles are adequately applied in practice to eliminate all the possible weak links in the external structure, and to ensure that no noise comes from plant or equipment which would produce any undesirable physical changes in the system.

The particular integrated insulation required will obviously depend on the level of external noise and the degree of disturbance to be eliminated. Various criteria can be put forward for the maximum levels of intruding noise permissible in acoustic design of systems. If the intruding noise is intermittent rather than continuous some account will normally have to be taken in

practice of the proportion of time for which these criteria levels may be exceeded. It is impossible to lay down hard and fast rules but the best attempt must be made within the bounds of practicability to reduce the intruding noise levels.

Absorption: All kinds of materials absorb a certain amount of composite wave elements (viz., physical agents). In noise control problems, it is necessary to predict as accurately as possible the amount by which surfaces, furnishings or materials affect noise fields by absorption. The absorption provided by many materials can be measured and tremendous practical data is available for comparison analysis. In noise control such data is essential to enable the best ways to be used to reduce the integrated noise.

The energy absorbed from noise of composite wave elements incident on a surface is most frequently converted into heat in the material, the remaining energy, transmitted or reflected, appearing as diffuse physical agents. The amount of energy absorbed obviously depends on the kind of material, but it also depends on other factors such as shape, angle of incidence and frequency of the incident noise field due to composite wave elements. The absorption coefficient, defined as the ratio of non-reflected to incident energy, can have values between 0 and 1. Absorbers can be used in situations in which noise fields are reverberant and composite wave elements are incident on the surface at all angles. For some materials the absorption does not vary appreciably with angle of incidence and data obtained with one kind of noise field may be more or less appropriate for other noise fields. For many kinds of materials and situations, however, there are differences and it cannot always be assumed that absorption charts obtained in one way can be used comprehensively.

Absorption mechanism play important role in classifying various types of absorbers.

Porous Absorbers: There are many materials commercially available for noise control fall come into porous absorbers. These are usually most suitable for general purpose treatment in noise reduction. The pores of such materials have to be interconnected so that fluid can flow right into the material and they absorb principally because of the viscosity of fluid. The noise waves cause motion of fluid relative to the adjacent material generating heat by friction and the constant compression and expansion of fluid in the pores with the resultant heating and cooling also contributes to the energy losses.

The factors which are responsible for absorption characteristics of such materials are the thickness, frequency, stiffness, porosity and structure factor. Some materials are essentially rigid and the absorption increases with frequency and thickness. The absolute values depend largely on the structure of the material. The principal disadvantage of such materials is the durability and appearance of the surface when used as a finishing material in wall. For example, rock wool and glass wool blankets are excellent absorbers but normally are unacceptable as a surface finish.

Panel Absorbers: Composite wave elements form noise field and which is acting over a thin impervious panel. Oscillations are set due to energy dissipated in the material and at the edges. The mass and stiffness of the panel on its supports and stiffness of the space behind govern a resonant frequency at which the absorption is at a maximum. Thin roofing sheets can be used over deep spaces to provide very low frequency absorption, and stiffer panels have their resonant frequency towards the low end of the spectrum. It is only at the low frequencies that this mechanism is effective since panels designed to resonate at high frequencies only provide absorption over a very narrow band of frequencies.

Resonant Absorbers: Any enclosed volume of fluid can act as a resonator. Such resonators are thought to have been used for noise control purposes since ancient times. Resonators tuned to a single frequency are rarely used in practice.

13.11. Energy Conversion in Engines and their Noise Fields

Energy balance of an engine: Energy or heat balance in an engine is an experimentally determined distribution of heat, which is added to the engine together with the fuel [41]. The energy balance is composed of input heat added to the engine together with the fuel in a definite time interval. The input heat is converted into: i) definite amount of useful work; ii) the amount of heat given to cooling fluid; iii) the amount of heat given to the lubricating oil; iv) the heat carried away from the engine by exhaust gases; v) the amount of heat that does not come out because of incomplete combustion; and vi) the remainder heat, which theoretically cannot be converted into useful work.

Thermal loads and thermal stress level: Thermal load is defined as the value of specific heat flux transferred from the working medium to the surface of a part. Transfer of heat from the working medium to the surface is affected in two ways: i) by radiation; and ii) convection. In engines, radiation is of particular importance, because mainly diffusion combustion occurs in them, which is accompanied by formation of soot which burns out subsequently. The soot content in the flame is the cause of its high degree of blackness and, therefore, of high emissive power of the flame. In practice, diesel flame temperature is much higher than the volume averaged thermodynamic temperature. High flame temperatures and degrees of blackness of flame are the cause of a high fraction of heat transferred by radiation. In practice, it is estimated that up to 45% or more heat is transferred by radiation. Radiation of tri-atomic gases also plays partial role in radiation transfer. The thermal stress level of separate parts of engine depends mainly on the disposition of the portion relative to the flame and is therefore not the same. In some engines, zones of the parts like the cylinder liner, cylinder head and piston are shielded by the piston body from the flame during intense radiation.

Heat transfer includes the convective heat exchange in the bulk of the charge and transfer of heat by conduction via the boundary layer of the charge. The heat transfer intensity is determined to a great extent by local conditions of mixture formation and heat liberation. Many case studies have revealed that the intensity and nature of motion of the charge in the cylinder and combustion chamber, formed during the intake, have not much effect on thermal loads in engine parts with variation of the motion of the charge initiated during combustion and the distribution of fuel over the combustion chamber volume. The fuel distribution

depends on the number and location of the fuel sprays and the dimensions and configuration of combustion chamber. These factors determine the local temperature of the charge. In summary, it has been revealed that the heat exchange in piston type engines is markedly nonstationary and the distribution of thermal loads in its parts is rather non-uniform. The nonstationary nature of the heat exchange depends on all the time-variable factors such as state variables of the charge, charge velocity, flame structure etc., which affect radiation and heat transfer. In such a heat exchange, the maximum specific heat flux may be several factors higher than time-averaged value. The greatest part of heat is transferred in the period of intense burning.

The type and dimensions of a combustion chamber have a significant effect on the non-uniformity of distribution of the thermal loads. Thus, in the piston type engines, the non-uniformity of thermal loads increases with a decrease of the relative diameter of the combustion chamber. In divided chamber engines, the non-uniformity of distribution of specific flux over the piston and cylinder head surface is, as a rule, more than in the open chamber engines. The level of thermal loads is to a great extent determined by the engine rating (brake power per litre). It is, as a rule, higher in two-stroke engines. Thermal loads appreciably increase with an increase in external loads (decrease in excess-air coefficient), speed, pressure and temperature of air at the engine intake. Also the injection (ignition) advance angle has a definite effect on the thermal loads. With an increase in this angle, the thermal loads grow owning to the increase in maximum pressure and temperature of the cycle. As time elapses, after a new or overhauled engine is put into service, the thermal fluxes transferred from the working medium to the parts decrease due to the accumulation of partial oxidation, cracking and polymerization products of lube-oil and fuel on the parts. Further, the transferred fluxes tend to stabilize. When the engine operates at variable conditions, the heat fluxes transferred from the working medium to the parts vary not only throughout every cycle but also from cycle to cycle. Abrupt variation of the time-averaged thermal loads in one cycle due to racing, loading, unloading, and stopping of engine is known as the heat shock. The nature and frequency of heat shocks affect the reliability of engine operation.

The term thermal stress level is used to express a set of phenomena related to the thermal state of engine parts. The thermal state of the parts has an effect on the strength characteristics of the material of which the parts are made, on the rate at which deposits appear on the parts, on the lubrication conditions of the parts, on friction, wear and stresses in the parts. Temperature stresses appear because of uneven distribution of temperature in the parts and also because the majority of parts do not enable the most heated portions to expand freely. Therefore, the thermal stress level depends on the distribution of temperature in the parts. Thermal stress is a function of the heat load, part design and its cooling conditions. The distribution of local thermal resistances depends on the design of parts. The cylinder head and piston are the most thermally stressed parts. The thermal state of the cylinder liner is also of importance, because it has an appreciable effect on the thermal state of the piston.

The methods of controlling the thermal stress level are, of course, dependent on the factors which determine the thermal stress. A definite permissible level of thermal loads corresponds to specific designs of parts, to the materials used and cooling conditions. Thermal loads can be reduced by the use of cooled constructions, which makes it possible to appreciably lower the temperature of piston. Heat insulating coatings aid in reducing the temperatures and temperature gradients of the parts. An appreciable reduction in the thermal stress level can be achieved through the use of rational design of the cooling system. Fluid cooling systems are, as a rule, more effective than air cooling systems.

Internal losses in engines: By internal losses is meant all kinds of friction losses, gas-exchange losses, auxiliary mechanisms (water, oil and fuel pumps, fan, generator, etc.) drive losses, ventilation losses due to the movement of engine parts in air-oil emulsion and atmosphere, the engine being running at high speeds, and also the compressor drive (two- and four-stroke engines supercharged by a drive supercharger) losses. In the divided-chamber engines, the hydro-dynamic losses due to the flow of the charge between the cavities of the combustion chamber are also classified as internal losses. The greater part (up to 80%) of the internal losses is frictional losses. Much of the friction losses are accounted for by the piston liner and piston rings-liner pairs. The bearing friction losses amount to 20% of the total internal losses. The following forces load the rubbing pairs: the forces of inertia, the forces exerted by gases and the elastic forces (of rings, springs). The estimation of time-averaged values of the forces acting on the parts is of significant importance for frictional losses. The time-averaged forces of inertia in absolute value are much higher than the time-averaged forces exerted by gases. Of the elastic forces, the elastic forces of the piston rings which are independent of the engine speed have the maximum effect on the frictional losses.

The following factors have a significant effect on frictional losses: i) thermal conditions of the engine affect the oil viscosity which essentially determines the forces of friction of complete fluid lubrication. When the engine runs at sub-zero temperatures the frictional losses increase and the effective characteristics of the engine deteriorate. The most appreciable decrease in the viscosity of motor oils occurs when the oil temperature mounts up to 60 °C; ii) Speed. Increasing the speed causes the forces of inertia and the relative rate of travel of parts to grow. At the same time the temperature slightly rises and the viscosity of the lubricating oil decreases. The friction forces of the complete fluid lubrication increase mostly due to the growth in the relative rate of travel of parts. The friction forces of the boundary lubrication grow due to the increase in the loads acting on the rubbing pairs. On the whole, the frictional losses appreciably increase with growing speed. Increasing the load leads to growing the gas forces, rising the temperature of parts and decreasing the viscosity of oil. Note that the speed, load, supercharging pressure, and compression ratio, on the one hand, and the design, dimensions of rubbing parts, quality and thermal conditions of lubrication, on

the other hand, should be correlated with the aim of ensuring reliable friction of the complete fluid lubrication. If these conditions and the engine operating rules are observed, then the frictional losses first decrease due to the running-in of parts and then stabilize.

Acoustic parameters of engines: The concentrations of a large number of mechanisms and machines equipped with engines are the cause of increased noise due to various physical agents at construction sites. These noise levels interfere in doing work and disturb the people in taking rest. Noise has an adverse effect first of all on the sensing organs of operators of machines equipped with engines, annoys people by getting on to the nerves, reduces the output and hinders in picking up of warning noise signals and human physiological responses. The standards as well as a correct estimation of the noise level at a work place serve as bases for the organization who handle the in-service engines. Generally the external and internal engines are normalized. There is definite tendency to lower the admissible values of the engine noise, which demands the development of new and innovative designs. The total noise level may serve as an integral characteristic of the quality of engine and of the culture of production and production processes. These separate characteristics of the engine noise are used as diagnostic parameters.

By engine noise is meant the acoustic radiation emitted by the engine when in operation. The engine noise is measured by the noise level and the noise spectrum. These are the engine noise characteristics at a point in the space. As a source of acoustic radiation, an engine is characterized by the emitted acoustic power, its spectrum and the directional diagram. Integrated noise spectrum is a very important characteristic of physical agents. The reaction of human sensing organs to physical agents of one kind of amplitude but of different frequencies is not the same. The integrated engine noise spectrum should show the distribution of the emission energy over the frequency range. Discrete components which are multiples of speed and number of cylinders can be plotted with a continuous region in integrated noise spectra. The integrated octave band spectra of all the powers of physical agents may serve as one principal characteristic of the machine noise.

Energy balance of acoustic parameters in an engine: Physical agents are known to appear due to-i) the interaction of vibrating body with the medium; ii) upon rapid liberation of heat in a finite volume of the medium; iii) during addition and outflow of a finite amount of substance to a definite finite region of the medium; iv) upon interaction of the flow of substance with a solid body. These physical processes occur precisely in time domain during accomplishment of the working cycle. Here, in all the cases of various physical agents, the acoustic emission will account all the oscillations and propagation of oscillations in this system and subsequent energy emission process in the surroundings. The combustion chamber walls deform during compression, combustion and expansion, which causes the outer walls of the engine to oscillate. The energy of wall oscillations in the form of dissipated energy is emitted to surroundings. During combustion, the heat added to working medium in the cylinder gives rise to the acoustic emission. The tilting moment will set the suspension-mounted engine into oscillations whose energy in the form of noise will partly dissipate to the surroundings. When in operation, shocks of contact parts may appear in the mechanisms and this produces additional noise. The work of units (fan, fuel feed pump, etc.) disposed on the engine also gives rise to one more noise term. During the exhaust, the substance inflows into the region adjacent to the exhaust manifold, which is liberated with a certain amount of energy. This gives rise to exhaust noise. Comparing the energy balance of acoustic parameters in an engine makes it possible to determine the most significant components of the engine noise, to indicate the causes of noise, to study the noise-formation process, and to find the most rational ways of reducing the engine noise.

Noise reduction in engines: The oscillating systems with distributed parameters of elasticity, inertia, and energy distribution are known to have an infinitely large series of frequencies of free oscillations. In reducing the engine noise, focus is to study the number of low-frequency free oscillations of parts, which fall in the range of the action of the force and its spectra. These set of low frequencies in combinations with the construction properties to absorb oscillations is called the amplitude-frequency characteristics of an engine part. The ability of the force to set up oscillations of the part on which it acts is determined first by the amplitude of the component force estimated by the coordinate of the spectrum. The larger the ordinate of spectrum, the greater is action of its force at a given frequency. Another important characteristic of the spectrum is its width. The wider the force spectrum, the greater will be its given force over a wider range of frequencies. This force can set up oscillations in the construction. Coincidence of frequencies of the amplitude-frequency characteristic of a part and the force spectrum is indicative of the development of oscillations of the part at the given frequency and thus, of the generation of noise. The release of energy of engine outer wall oscillations to the surroundings depends on the radiation resistance. In the general case, this results in active and reactive components of the radiation resistance. The active component reveals what part of the energy of oscillations is radiated as noise to the surroundings, and the reactive component gives the fraction of energy which the medium and oscillating body exchange during the period of oscillations.

The intake and exhaust noise are traditionally reduced by installing silencers. The intake silencer is usually designed integral with the air cleaner. Note that the air cleaner lowers the intake noise to a definite extent. However, modern engines are often equipped with additional mufflers at the intake. These are, as a rule, resonance chambers. The ability of a resonance chamber to muffle noise depends on its volume, the length and the area of the section of the connecting neck. Theoretically, the muffling would be the maximum only at a one resonance frequency which depends on the combination of these parameters. If an integrated noise absorbing material is disposed in this resonance chamber, then the muffling at the resonance frequency will slightly decrease but the frequency range of muffling will widen. For a better muffling, an exhaust system consisting of an exhaust silencer can

be designed. As a rule, the exhaust system designed for an engine to be used in an automobile or a tractor cannot be utilized in road building machines due to design complexities.

Noise can be reduced by applying the methods that involve vibro and noise proofing, vibro and noise absorption muffling. Vibro-absorption methods are widely used in the motor engine industry. The vibro-absorption effect resides in making use of internal friction of structural materials to absorb the vibration energy. As a result, the vibration energy converted into heat energy. This property of the construction can be enhanced, i.e. special devices –vibration energy absorbers – can be created and the engine is equipped with them. It is good experience to choose structural materials for engine parts with due regard for the internal friction. The characteristic of dissipative properties of a material, the coefficient of losses or any other characteristic should be taken into consideration along with the strength and wear-resistance characteristics. Use is also made of the oscillation or sound isolation properties which hinder the vibration energy propagation over the construction or in the engines.

Internal combustion engines may be equipped with special noise proof devices-capsules. The capsule does not contact the engine surface. Here, the engine supports are made in such a way that they exhibit high vibration oscillation properties. In designing other devices such as noise control units, similar requirements are complied with to prevent vibrations from appearing at the outer surface of the capsule, otherwise the useful effect will be lost. The noise proof devices are shields which partly block the way for a flux of a physical agent. Here it is possible to act on one or several components of energy balance of acoustic parameters on the surface. On the whole, it should be said that the maximum noise reduction can be obtained if vibro and noise proofing devices are used in combination with the vibro and noise absorbing constructions.

13.12. Noise Pollution & its Control

Our earth environment is composed of air, water and land and these are known as atmosphere, hydrosphere and lithosphere respectively. All these spheres make the biosphere. In the biosphere, apart from human beings, plants, animals, birds, fishes, insects and microorganisms also exist. The atmosphere provides oxygen, while the hydrosphere and lithosphere are responsible for water, food and space. Whenever a change, physical or chemical, occurs in the biosphere, all living beings get affected. This change is termed as pollution and the agents that institute these changes are called pollutants. Noise pollution is caused by unwanted intensity of these agents; it may be due to pollutants of air, water, light, sound, heat, radiation, fire and electromagnetic waves. The major sources of noise pollution are: i) industry and machinery; ii) transportation (surface, air and sea); iii) community activities (construction works, events, entertainment, etc.).

The main menace of noise is the transport systems. Noise from road transport, particularly heavy diesel vehicles and motor bicycles, can be a nuisance. The use of larger and faster jet aircraft is a menace if the airfield happens to be in one's neighbourhood. It causes significant damage to property besides giving freight to humans and animals. It can cause mental anguish too.

Noise pollution control is achieved by: i) reduction of noise at the source. For example, noise from a fan can be reduced by increasing the number of blades or by decreasing the rotational speed without reduction of airflow; ii) noisy machinery or equipment may be covered with insulating material; iii) protection devices like breathing masks, scarfs, mufflers, earmuffs may be used to prevent body from heat, air pollutants and excessive sound; iv) Acoustic zones should be created to prevent propagation of noise; v) trees can be planted in and around the noise sources to attenuate noise.

Noise pollution control by road transport can be reduced by using transport services like metro, bus, electric or CNG vehicles and by preventing honking of horns. The sides of the roads or highways could be lined with trees, which act as a buffer and reduce the noise pollution. New traffic regulations enforcing lower speed limits can prove beneficial. Old vehicles which produce lot of noise pollution should be taken off the road. Noise pollution at community level can be curbed by imposing a ban on: i) disposal of solid & kitchen wastes; ii) a ban on blaring loudspeakers and by reducing their pitch; iii) ban of sale and use of crackers during festivals, and processions.

Of all the measures of noise control the elimination of noise at the source, i.e. directly in the apparatus or industrial process mechanism is most effective. Some of the measures for the suppression of noise at the source are: i) substitution of non-percussion processes and tools for percussive ones, e.g., use of hydraulic drives instead of cam or eccentric mechanisms, welding instead of impact riveting and straightening instead of forge-rolling, etc.; ii) use of rotational, preferably uniform, motion instead of reciprocating motion; iv) replacement of spur and chain gearing with Y-belt and simple gear-belt transmissions; and v) replacement of spur chain pinions with helical and shevron gears.

Integrated noise proofing includes construction of barrier structures, such as walls or partitions, to safeguard the workers from external noise. Integrated noise proofing utilizes principle of reflection of composite wave elements, i.e., the greater part of the energy incident on a surface is reflected and only its smallest part penetrates through it. Ideally a noise proof structure should not let noise into an enclosure it safeguards. For the case of sound energy, it penetrates through obstacles using their own vibrations. In other words, the obstacle itself becomes the source of noise and radiates sound energy into the room, it is supposed to protect. Therefore, the more heavy (massive) the barrier structure, the more sound-proofing is achieved. As the acoustic resistance of the surface is determined by its acoustic inertance it is more sound proof to high frequency waves than to waves of low frequency.

Noise absorption is the process by which energy of composite wave elements is diminished in passing through a medium or striking a surface. Another benefit from noise absorption is the possibility of easy checking of human senses, due to operation of industrial processes, where direct composite wave elements succeed noise from every apparatus. Ceilings and the upper portions of walls above the floor level should be lagged with noise-absorbing material.

Aerodynamic noises are sharp pulses of velocity and pressure of air or gas flowing in conduits, occurring during the operation of pneumatic machines, engines, compressors, turbo-blowers, fans, ejectors, etc. To abate such noises at the source is practically impossible so they have to be suppressed or silenced. This aim is achieved by using various types of silencers and exhaust mufflers which deaden the noise by weakening the pulsations of pressure in a flow of air without causing resistance to the flow at the outlet.

13.13. Fire Alarm Systems

The objective of fire alarm signalling system is to detect and warn people in due time and inform the public fire station about the place of fire. In a building, an alarm system should be available to warn everybody in case of fire. If the alarm is not given automatically this can be done by installing alarm bells, whistles, or sirens in different parts of the building, and push-button or handles in all workrooms to operate the alarm, if necessary. The automatic fire alarm system comprises of fire-detectors (transducers), a power source, and a fire alarm station connected electrically. Electric fire-signalling system can be of two types: the radial line and the closed loop system. In a radial type fire alarm system each transducer is connected to the receiving station by a separate two wire (out and in going) line. The receiving station is designed similarly to a telephone switching center to receive all the signals from fire detectors. The system has proved reliable and effective as it provides simultaneous reception of signals from all lines.

Fire alarm warning devices, called fire-detectors, are in fact transducers that convert the physical quantities characterizing fire into electric quantities, i.e., signals which are collected at the receiving station. The signal can be initiated manually (push-button) or automatically as soon as the parameters of the environment change due to fire (temperature, flame, smoke). Fire alarms can be heat, smoke and light detectors, or detectors that use more than one parameter, such as heat-smoke sensitive detectors.

According to the principle, the fire-detectors are designed to operate on alarm and a distinction is made between the maximal, differential and maximal-differential types of detectors. The maximal-type fire detectors are designed to operate as the temperature in the room reaches the pre-set value (+60 or + 80 °C, regardless of the rate at which the temperature rises). The differential type detectors operate as the rate of temperature rise reaches a pre-set value. There are however quick-response heat detectors that ensure as instantaneous warning at the very onset of fire or even when the environmental changes preceding fire such as flashes, flames, sparks or smoke are low-key. This is ensured by using photo-cells, ionizing chambers, semiconductors, thermocouples, etc.

Some heat detectors are thermocouple devices, the operation of which depends on the difference in temperature between the junctions of two dissimilar metal conductors. In a closed circuit, this temperature difference generates an emf, thermos-electromotive force, which provides a thermoelectric current. The magnitude of the thermos-emf depends on the nature of metals and difference in temperature. A combination of two dissimilar metals having a junction which can be maintained at the temperature which it is desired to measure in terms of the thermoelectric current produced is known as a thermocouple. An individual thermocouple produces a small thermos-emf of several millivolts. To increase the thermos-emf a number of thermocouples are connected in series or in parallel or both to form a thermopile.

Some heat fire detectors use a temperature-sensitive resistor with a negative temperature coefficient. Heating increases its resistance due to increased number of free charge carriers. Resistors are normally fabricated from oxides, sulfides, and metal carbides (iron, nickel, manganese, titanium). Smoke detectors employ photo-cells or ionizing chambers or differential photo-relays. Ionizing chambers using alpha rays proved most popular. The ionizing power of alpha particles is higher than of beta or gamma rays and less harmful to the human body. Light detectors are designed to respond to different regions of the open flame light spectrum. Most popular are light detectors which respond to the ultra-violet region of the optical spectre and which may include thermos-resistors, photo-cells, photo-resistors or photo-counters, etc.

Fire alarm receiving stations or centres are designed to monitor and detect faults in the circuits, receive alarm signals, and to actuate automatic fire-fighting devices and installations. Radial type optical fire alarm station is designed for use in industry. It consists of a receiving unit with a power source and 10 radial terminals connected to it electrically. The number of fire detectors on each terminal is not limited. The centre receives alarm signals, checks the circuits for faults, warns the fire-brigade and operates automatic fire-fighting.

Radioisotope (fire-guard) stations employ smoke detectors, light and audible fire alarm indicators and are intended for use in industry. One such station is capable of initiating a fire alarm, controlling automatic fire-fighting and guarding premises against trespassing by controlling the integrity of the interlocking loops and giving audible and light signals as soon as the contacts of the lops are closed. Automatic fire alarm-guarding stations are designed to detect smoke, heat and open flames, indicate places of fire and warn the fire-brigade by light and audible signals. Such stations can also operate automatic fire-fighting and safeguard the premises.

Photoelectric fire alarm-guarding systems intended to receive and register fire-alarm signals coming from smoke detectors. It operates automatic fire-fighting and an audible alarm. The system ensures that a warning ‘Attention’ is given after one fire-detector operates, then it starts an ‘Alarm’ and automatic fire-fighting after two or more fire-detectors have operated. The system monitors and checks all the circuits and detectors, received the signals and transmits the information to the fire station. Concentrators of the fire-guard signalling system safeguard the building premises against trespassing and fire. Using one receiving station, the system serves the dual purpose of giving fire-alarm warning and safeguard signalling.

13.14. Noise Instrumentation

In order to measure intensities of physical agents, various types of sensors, transducers and meters are required. The sensing and transducing capability depends on the signal strength and its conversion into electrical signal for further processing, conditioning and data storage. Frequency analysis of the signal is required for its noise characterization. Much discussion on instruments for noise measurement is given in Purkis [42, 43].

Microphones: The prime necessity in any instrument for measurement is a microphone capable of reproducing as closely as possible in electrical terms the characteristics of the sound wave to which it is exposed. Although in some cases only relative measurement is required, such as when measuring the reduction afforded by a particular construction or device, in most cases an absolute measurement of pressure is required, i.e. the sensitivity or electrical output for a given sound field should be known, and the way in which the sensitivity varies with frequency and with angle for fields in which the sound is coming from a particular direction. Non-linear distortion will occur at sufficiently high sound pressure. The characteristics of microphones can be best categorized into electrical and sound. For electrical parameters, unit of *oncisol* should be used, whereas for their sound related parameters *oncibel* should be used. For louder noises, specially designed insensitive microphones may be required, since normal microphones may suffer actual physical damage.

Noise level indicators: The most common form in which intensity of a physical agent can be measured can be noise level meter, which can be built into the measuring amplifier. Indication of the r.m.s. value of the voltage or pressure is most often required. Thermocouples can be used to give direct reading but their use in practice is limited by their slow response and inability to withstand overloading without damage.

The noise measurements can be stored in a noise data recorder with computer storage. All the indicating devices give relative readings and for absoluter readings (of flux, pressure, intensity of a physical agent for example), a calibration signal, either of known level or electrical in relation to the sensitivity of the sensor used, is necessary.

Frequency analysis: In noise control, a single reading corresponding to the overall pressure, flux, intensity of a noise source, at a particular instant of time, is of limited use. Many practical measures for the control of noise by insulation or absorption are frequency dependent, i.e., have more or less effects for some frequencies in comparison with other frequencies. In subjective assessment, the spectral constant is used. Thus a frequency analysis of many noises is essential. The degree of analysis required varies with the kind of spectrum likely to be obtained and the purpose for which it is required. Although when presented with a completely unknown noise, it is only possible to find out whether the spectrum is highly irregular and has discrete components by a narrow band analysis. However, it is only experienced by practice that when a broad band and when a narrow band analysis is desirable.

Most of the filters used for broadband analysis are passive, i.e. consist only of arrangements of inductances, capacitances or resistances with no additional amplification and their specification in terms of the centre frequencies and response outside the pass band are a standard practice internationally.

Vibration measurement: Vibration is measured by instruments based on the mechanical and electric methods of measurement. Mechanical measuring devices, due to gaps and lag of the indicating and recording mechanism, give rather accurate readings only at relatively large amplitudes (more than 0.05 mm) and at low frequencies (up to 30 Hz). Electrical instruments are superior to and more precise than the mechanical devices. They permit measurement over a wide range of frequencies for vibrations of a large and low intensity; vibrograms are obtainable at a considerable distance from the vibration body, which is both safe and convenient. Vibration meters are used to measure vibration of surfaces. They are connected through transition adapters to the vibration transducers with the output characteristics corresponding to those of the sound level meter to indicate levels of the oscillatory velocity, instead of sound levels of pressure.

Integrated noise measuring instruments: Various instruments are used to measure noise levels among which noise meter is the commonest. This can be any form of meter which allows the operator to measure either objectively or subjectively. The measurement of loudness can be measured objectively. The principle of instrumentation is conversion of sound energy into electric energy and consists of non-directional piezo-electric microphone, an amplifier, connective filters, a detector, and an indicator. The overall circuit design is such that the meter's characteristics should approximate the response of human senses. Vibration noise meter (vibrometer) is an apparatus for the measurement in the octave band frequency or generally of the root mean square pressure levels of noise.

The levels of ultrasonic radiation can be measured by ultrasonic detectors and devices for the measurement and detection of ultrasonic waves. Such devices can be mechanical, electrical, thermal, or optical in nature [44].

To estimate electromagnetic radiation intensity in close vicinity of HF sources, measurements of electric and magnetic field strength are made separately. It is necessary because an electromagnetic field in the induction zone is a totality of independent electric and magnetic fields. The electric or magnetic field strength at a given location results from the passage of radio waves. In the case of a sinusoidal wave, the root-mean-square (RMS) value is commonly used. Unless otherwise stated, it is taken in the direction of maximum.

The principal instrument for measuring the intensity of electromagnetic fields is the radio –field strength meter, a devise consisting of an amplifier, dipole and frame antennas, and a voltage divider. The devise is calibrated using the range of the working frequencies to give readings of the effective field intensities from 5 to 1000 V/m for the electric field component, and from 0.5 to 300 A/m for the magnetic field component. The error of measurement does not exceed 20%. The antenna is placed in the magnetic field to be measured, and its position is varied until the scale gives a maximum reading.

Measuring noises from physical agents require sensing and transduction of certain parameters. Light meters can be used to measure illumination levels. Infrared thermometers can be used to measure thermographic imaging with temperature measurements. Solar intensity measurements can be made from a pyranometer. There are many heat flux sensors available to measure intensity of a heat source. A simple voltmeter can be used for measuring operating voltages of electrical equipment. Other electrical instruments are wattmeter and power factor meter for measuring power consumption and power factor. Combustion analyzer is used to measure combustion efficiency of furnaces, boilers in order to estimate energy losses and draw Sankey diagrams for exergy analysis. Ultrasonic leak detector is used for compressed air leak detection. These are electronic ultrasonic receivers that are tuned very precisely to the frequency of the hissing sound of an air leak. Airflow can be measured by a velometer, an anemometer, or an airflow hood.

Variation with time: Many of the noises requiring measurement are non-steady, and although it is possible, perhaps with the aid of measurements, to make an analysis of such noises and to obtain a response of noise level variation with respect to time, either for the full frequency range or in particular bands. Counters can be used to measure the noise band intervals. For example noise level ranges of 66 *oncibel* (oB), with a counter each placed at 6 oB range would require 11 counters to cover the entire 66 oB range. A fast counting rate, such as 11 per second, will be satisfactory for most kinds of noise where the intensity level is not varying too quickly with time, or alternatively a slower rate with some means of adding a number of very short intervals can also be used in the analysis.

13.15. Operation of Power Systems

Operating conditions and characteristics: The principal operating objective for any plant equipment and devices is to satisfy the output demand according to the load curve and ensure the highest service life and highest efficiencies with minimum noise generation in the plant. The load curve of plants is mostly non-uniform which is determined by variations of services and goods as received by the consumers. There are daily, weekly and seasonal load curves. Any plant equipment may operate in the base, semi-peak or peak modes. In a base regime, the load is essentially constant, though certain variations are allowed in this regime and even shut-down to reserve on non-working days. In semi-peak regime, the equipment is shut down to reserve in the off peak and weekends. In peak regime, the equipment is in operation only during peaks of the load.

The following principal characteristics are adopted to describe the operating conditions of any equipment: i) the net efficiency of the equipment at rated load and the average efficiency for a particular operating period; ii) the operation factor which is essentially the ratio of the actual operating time of the equipment to the calendar time of the period considered; iii) the availability factor, i.e., the ratio of the total time during which the equipment was in operation and reserve to the calendar time; iv) the capacity factor which is the ratio of the generated product to the probable product generation during the calendar time if the equipment operated at the rated product generating capacity; and v) the average and maximum time of a campaign or, using terminology from the reliability theory, the operating time to failure (a failure is understood as an event such as noise generation that reduces the output as well as disturbs the operating ability of any equipment).

Equipment operating in the base regimes should have the highest values of the indicated characteristics. It should be clear that such high indices cannot be attained by the equipment operating in semi-peak and peak regimes.

Since monobloc units gradually replace older equipment, some of them, including those are employed increasingly for supplying the variable portion of the load curve. In that respect, the manoeuverability of equipment is a characteristic that grows in importance. The concept of manoeuverability of monobloc units includes: the characteristics which determine the range of operating loads of a unit; the startup-shutdown characteristics; the dynamic properties; the characteristics at sudden sheddings and surges of the load. The conditions of operation of equipment at various loads are characterized by the load control range and the range of allowable loads. In the load control range, the automatic control system of equipment should quickly respond to load variations without interference of the operating personnel. In this range, it is not allowed to change the composition of the auxiliary equipment and the adjustment of automatic regulators. The concept of the allowable load range includes, in addition the region of loads from the lower limit of the control range to the lowest capacity at which the equipment can operate steadily. In this range, the load varies relatively slowly according to the load-dispatching curve and it allowed varying the adjustment of the auxiliary equipment to stabilize the process load.

Regimes of operation at any load with insignificant variations of output parameters are called steady. In contrast, regimes characterized by load variations and deviations of output parameters due to internal and external disturbances are called unsteady. Internal disturbances of an operating regime are caused by variations of one or more equipment inlet parameters. External disturbances are caused by variations in the external conditions. An important characteristic of equipment is its ability to change the load quickly, which is called the acceleration characteristic. It is determined mainly by the dynamic properties of equipment, i.e. by its response to regime disturbances.

The characteristic of manoeuverability of the equipment are determined largely by startup-shutdown regimes. The principal types of equipment shutdown are as follows: i) shutdown to reserve or for some types of repair work; ii) shutdown for repairs of whole plant; and iii) emergency shutdown; it is carried out by the prescribed procedure which is determined by the cause of shutdown and the possibility of restarting.

To illustrate the process of operation of power systems, the case of boiler combustion of thermal power plant is mentioned herewith [45].

Power systems operation under steady conditions: In general, equipment operation is carried out according to a regime chart which gives the principal characteristics of the regime

obtained by the equipment tests. Modern monobloc units are equipped with computer panels which make it possible to determine the current or average efficiency characteristics and the process stream characteristics. This information is used for the optimization of equipment operation. In addition, a number of automatic regulators maintain the process parameters at optimal values. For these devices the regime chart gives the extreme variations of the variables, above which the personnel should correct the operation of the controllers or change to manual operation. Steady regimes of the equipment operation may be different depending on the load. Variations of a particular process parameter or regime characteristics as a function of load are described by what is called the static characteristics.

For the case of thermal power plant, controlling the efficiency of combustion process consists in maintaining the optimal excess air ratio in the furnace and in distributing the airflow between the burners in accordance with fuel distribution. Uniform distribution of fuel and air between the boiler burners is favourable for decreasing the maximum temperatures of water-wall tubes, decreasing temperature stratification of flue gases at the furnace outlet, and preventing slagging of the heating surfaces. The temperatures conditions of the heating surfaces are also impaired if some tubes have leaks or bursts. Tube-burst indicators of the acoustic type based on sound sensing are finding increasing use for detecting bursts and leaks in boiler tubes. If the temperature behind particular heating surfaces has increased above the allowable level or if a tube burst has been detected, the boiler is shut down.

With an increase of load, the temperature of flue gases at the furnace outlet increases, resulting in a greater volume of combustion products and higher temperature gradients in the convecting heating surfaces whose heat absorption thus increases with this result. As has been established by boiler tests and by experience of boiler operation, the lower limit of the load control range is: i) 40-50% of the rated load in the combustion of fuel oils, gas and high-volatile coals; ii) 50-60% for lean solid fuels; and iii) 60-75% for slagging-bottom boilers. The lowest load for stable operation of a boiler is usually 30-40% of the rated load.

Power systems operation under unsteady conditions: Specific processes occurring in drum-type and once through boilers are determined by the differences between these boiler types: the position of the boundaries of the steam-generating portion, its filling mass and capacity of heat accumulation. For a drum-type boiler, a change in load is done by changing simultaneously the fuel consumption and the capacity of blowers, with subsequent variation of feed water flow rate which is controlled by the water level in the drum, with leading pulses from variations of the steam and feed water flow rates. The load of a once-through boiler can be varied by similar procedures, with the sole difference being that feed water flow rate is controlled by the fluid temperature in an intermediate portion of the path. A different control procedure is often employed in the automatic control systems of once-through boilers: first the feed water flow rate is varied according to the load and the flow rates of fuel and air are controlled in proportion to it and corrected by the fluid temperature in the intermediate portion of the path. In monobloc units, two versions of load variations are employed. For planned load variations, the load of the turbine is changed by varying the steam-generating capacity of the boiler so as to maintain a constant pressure of live steam. When the unit operates in a regime controlled by the power controller of the turbine, the load of the turbo-generator is changed first and the boiler load is then corrected.

Starting-up circuits: The starting-up circuit includes special starting devices, which are used only at start-up, shut-down and during load shedding of the equipment. Taking example of a boiler unit in thermal power plant, these starting-up circuits include devices for discharging the steam from the boiler into the atmosphere before the boiler is connected to the steam main header. In Monobloc units, the start-up and shut-down regimes of the boiler and turbine are interrelated more closely and are much more complicated. In this connection, starting-up circuits, operating instructions and automatic control systems for monobloc units have been symbolized. They are based on a standard procedure of starting up a unit by 'sliding' regime, i.e. by gradually increasing the flow rates, pressures and temperatures of live and reheated steam. This procedure creates the most favourable conditions for starting the turbine and decreases starting-up losses. The procedure requires no special devices for its realization in monobloc units with drum type boilers. For the quick starting-up of a boiler, the starting circuit has a pipeline for steam discharge from the drum into the atmosphere, which is in use during the period when pressure rises roughly to 0.6 MPa. The sliding regime of starting-up of a once-through boiler cannot be performed without special devices, since from the requirements of reliable temperature conditions and hydrodynamics of steam generating water walls, the starting-up flow rate of feed water must be less than a specified value and the fluid pressure in the water walls must be maintained at a level close to the nominal pressure. For a sliding starting-up regime, a once-through boiler is provided with an internal starting-up unit which comprises an internal gate valve, an internal separator, and pipelines with throttle valves for the supply of working fluid to the internal separator, drainage of moisture, and steam removal from the internal separator. The starting-up circuit for any type of boiler has a bypass line around the turbine which connects the steam main directly with the condenser. A startup-shutdown device in this line comprises a closing throttle valve, a noise absorber for sound waves, and a spray-type steam attemperator. Live steam is discharged through the turbine bypass line in the following cases: i) at starting-up of the unit, for a period when steam temperature is raised to the level determined by the thermal conditions at the turbine inlet; ii) at shut-down, for the period during which steam is discharged from the boiler and the latter is cooled; and iii) at load shedding, for the period when the turbine is idle-running or supplies only the auxiliary load of the unit. Starting-up circuits have special devices to maintain the temperature of live and reheated steam. A characteristic feature of drum-type boilers is that their steam-generating capacity during start-up lags substantially behind the increasing fuel consumption, so that the

temperature of tubes of super-heater rises substantially and may exceed the allowable limit. Besides, it is essential to maintain the specific schedule of live steam temperature variation before the turbine. On the other hand, the condensate-spraying system of the main circuit can operate only at boiler loads not less than 30% of the rated value. For this reason, there is provision of a starting-up spray system with atomizing nozzles arranged in the condensate-spray diffusers.

In most cases, the starting regime of a monobloc unit should ensure a reduced reheated steam temperature level and further control of this temperature according to the specified schedule. In view of the dynamic properties of the re-heater and some other factors, the main means of control can be used only beginning from loads 25-30% of the rated value. A method for starting control that is widely used in all types of monobloc units is steam bypassing in the reheating system. Part of the steam is passed from the cold reheat line into the hot reheat line thus lowering the temperature gradient, and therefore, the heat absorption in the re-heater. As steam from the cold reheat line is mixed with that in the hot reheat line, the temperature of the reheated steam can be controlled at the lower level. In once-through boiler units, a starting spray is additionally provided in the steam main, which is used to control the temperature of reheated steam upon closing the steam bypass. Feed water to the starting spray is supplied at a pressure of more than 7 MPa from the intermediate stage of the feed pump, so that there is no need to control the water pressure in the spray line. In some cases, only one means of reheated steam temperature control is employed. It is possible to employ steam bypassing only, if it has been established by boiler tests that the re-heater has reliable temperature conditions at the maximum load at which the steam bypass will be used. The allowable conditions for starting spray are determined by the minimal flow rate and pressure of steam at which moisture will be evaporated completely in the steam flow and will not precipitate on the internal surface of the steam main header, especially in the nearest bend after the spray.

Power system shut-down and load shedding: Before normal shutdown, the boiler of a monobloc unit should be unloaded, i.e. its load should be shedded. When shutting down for short time (for instance, overnight), the hot state of the plant is usually retained wherever possible, so that load shedding is limited (for the most part not less than 50% of the rated load). When shutting down for a longer time, it is advisable to utilize the accumulating capacity of the boiler as much as possible. This is favoured by the deep unloading of the unit with sliding (variable) pressure of live steam. In addition to this, upon flame extinction in the boiler furnace, the turbo-generator is left for a certain time connected to the system so as to work off the accumulated steam in the boiler. The load shedding conditions in a drum type boiler are determined by the rate at which the saturation temperature in the drum decreases; this should not exceed 1.5-2 °C/min. The rate at which load shedding is occurred in the once through boiler by the action of its automatic control system can be determined by the dynamic properties of the boiler.

A boiler is shut down by interrupting the fuel supply to the burners and the supply of feed water. In drum-type boilers, the drum is first filled with water to the upper working level. Upon flame extinction in the furnace, the draft fans remain in operation for 10-15 minutes to prevent the accumulation of an explosive mixture in the gas-air path, which is thus ventilated. During the standstill, measures are taken to prevent the condensation of steam in super-heater tubes, since this might involve complications in subsequent firing of the boiler and decrease the boiler reliability. In view of this, as a once-through boiler is shut down to reserve and the feed water supply to it is discontinued, the water path up to the internal gate valve is closed (by closing the feed water valve, and the fittings of the internal separator) and steam is discharged from the super-heater (via the start-up and shut down device) and re-heater (via the discharge line from the steam mains) into the condenser. When shutting down a drum type boiler, the pressure in it is reduced gradually (by steam discharging through the start-up shut down device of the de-super heater), so as to maintain the saturation temperature below the temperature of flue gases in the super heater, i.e. by forming conditions preventing the condensation of live steam in the super heater. The drum is periodically fed with water. The re-heater is freed from steam in the same way as in a once-through boiler.

When the cold shut-down of a monobloc unit must be performed, requiring that all boiler equipment be cooled, load shedding is carried with the gradual reduction of the pressure of live steam. In a once-through boiler, the internal gate valve is closed at a load roughly 60% of the rated value and the pressure in the path up to this valve is then maintained constant. In both types of boilers, the temperature of live steam is lowered together with its pressure at a rate allowed by thermal stresses in metallic elements of the boiler, pipelines and turbine. At the same time, the temperature of reheated steam is lowered. The super heater and re-heater can thus be cooled roughly to 300 °C.

In emergency situations, a boiler is shut down with ‘pressure retention’. As the furnace is extinguished, the feed water control valve and the team outlet valve are closed, the turbo-generator is switched off, and the start-up shutdown device is kept closed if the pressure of the live steam is below the actuating pressure of the boiler safety valves. In this method, the pressure and temperature of fluid along the boiler path are first held at the same values as in normal operation. After that they diminish slowly owing to heat transfer to the surroundings and loose fittings of the water steam path. If the trouble is eliminated quickly, this method of shutting down makes it possible to start the boiler from the state of hot reserve. Otherwise, shutting down is continued and steam from the super-heater is discharged.

Emergency shut-down of a boiler must be carried out immediately in the following situations: i) when the water level in the drum passes beyond the safety marks or the supply of feed water to one of the flows in a once-through boiler is interrupted for more than 30 secs; ii) on failure of the water level gauges in a drum type boiler or of the feed water flow meters in a once-through boiler; iii) when there is no flow of steam through the re-heater; iv) when the pressure has risen intolerably (in once-through boilers, a pressure drop is also dangerous); v) on rupture

of tubes in the water steam path; vi) on flame extinction in the furnace, explosions in the furnace or gas ducts, or inflammation of combustible deposits in convective heating surfaces; vii) on an intolerable reduction of the pressure of gas or fuel oil the control valve; viii) when there is no voltage at remote-control devices and measuring and control instrumentation; and ix) on switching off of the turbine (in monobloc units) or of some of the auxiliary equipment (draft fans, blowers, etc.). In all such cases, a delayed shut-down can aggravate the situation and cause serious damage, so that the boiler operator is instructed to stop the boiler without waiting for permission from the management. An emergency shut down of a boiler is affected by the protection system on receiving the signal as per appearance of some or other emergency situation.

Firing system for thermal power generation: Firing system operation consists of preparatory procedures, firing, and raising the load to its specified rated capacity. The first preparatory stage includes assembling the water-steam, fuel, and gas-air paths, preparing all mechanisms and systems, creating vacuum in the turbine condenser, pre-starting de-aeration of feed water, etc. A drum type boiler is filled with water as required. The water level in the drum should be somewhat below the normal mark so as to allow swelling. A once-through boiler is to be filled with water in all firing regimes, except for firing from the state of hot reserve. As water is fed into the boiler, it displaces air from the system (provided that the air pressure is not excessive). In a once-through boiler, the feed water flow rate is adjusted at the starting-up level and the water pressure is raised to the working value by closing the throttle valve, In boiler firing from the hot state, first a reduced flow rate of water is established (10-15% of the nominal value), which makes it possible to slowly cool the boiler path up to the internal gate valve and internal separator. The starting-up flow rate of water is set in upon raising the water pressure before the internal gate valve. Water from the internal separator is discharged into the start-up expander and further into the circulation conduit. The start-up and shutdown device is opened to create vacuum in the super heater. This procedure is also carried out in a drum type boiler if there is no excessive pressure in it; this ensures a slower rise in the saturation temperature of the drum during firing. In cases when the start-up and shutdown device is initially closed, its opening is performed only upon firing the furnace, so as to maintain a constant pressure of live steam that has been generated to that moment.

During boiler standstills, moisture may accumulate in some stages of the super heater, which can be prevented through certain measures. Besides, un-tightness of the internal gate valve and throttle valve in a once-through boiler may lead to moisture accumulation in the pipeline and the first heating surface behind the internal gate valve. This moisture can be pushed out into the hot headers of the boiler during firing and cause their cracking. In a drum-type boiler, this may result in a quicker pressure rise in the drum at the initial stage of firing, which in turn diminishes the allowable heating intensity in the furnace. By opening the start-up shutdown device, the super heater is connected with the condenser which promotes the evaporation of moisture from the super heater tubes during boiler firing.

After switching on the draft fans, ventilating the gas air path, and preparing fuel supply mechanisms, the regime of burner firing is carried out. It is advisable to carry out the boiler firing by igniting as many burners as possible with the least fuel flow rate through each of them; this is essential for the uniform heating of the water walls around the furnace periphery and for minimizing local thermal stresses; in drum type boilers, this measure is additionally essential for the development of circulation in all water walls. In boiler starting from a cold or warm state, the initial flow rate of the fuel is controlled at a level of 12-15% of the nominal value. In a drum type boiler, this ensures the quick development of circulation in water walls and, on the other hand, the rate of pressure rise in the drum is within tolerable limits. In starting from the hot state, the fuel flow rate in the initial period is established at 20% of the nominal value or, if there is steam flow through the super heater, is increased additionally so as to attain the required temperature of live and reheated steam.

As the initial flow rate of fuel has been established in a once-through boiler, the flow rate of the feed water and the pressure of fluid before the internal gate valve are maintained constant. As the pressure of fluid in the start-up expander has risen up to 0.4-0.5 MPa, steam from the expander is removed to the deaerator and, when a sufficient quantity of drain water has been accumulated, the cycle is closed. In boilers with boiling-type economizers, the regime with periodic water feeding and a low constant flow rate of water can in some cases lead to a substantial maldistribution of heat along with hydraulic maldistribution. Some water-distributing tubes may supply the fluid of a higher enthalpy (up to super-heated steam) to the drum. This is prevented by maintaining the specified temperature of fluid in an intermediate section and at the economizer outlet by properly controlling the flow rate of water and by increasing the rate of blowdown on an increase in the water level in the drum. As the initial flow rate of fuel has been established, it is possible to raise the flow rate and parameters of live steam in a drum type boiler or the temperature of fluid before the internal gate valve, in a once-through boiler. The temperature of fluid determines its moisture content on entry to the internal separator.

As the super-heater is connected to circuit, as a result the temperature of tube metal in the heating zone decreases. At the same time, the temperature of steam at the boiler outlet gradually increases, since the heat transfer coefficient increases with the flow rate of steam. As the steam flow rate through the super heater increases, the steam mains are gradually heated. Steam is discharged from them through the start-up shutdown device and the drainage of the dead-end sections. Heating is usually carried out until the temperature of steam before the high pressure cylinder of the turbine is roughly 100 °C higher than the temperature at the steam inlet. In units provided with a pressure reducing de-super-heater, the re-heater system is heated by supplying live steam to the cold reheat pipeline and discharging it from the hot reheat pipeline into the condenser. This heating is only started when the temperature of steam before the de-super-heater exceeds the temperature at the discharge end of the high-pressure cylinder of the turbine, so as to avoid cooling of that

section. The end of heating of the hot reheat pipeline is determined by the temperature of steam before the intermediate-pressure cylinder of the turbine which exceeds the temperature at the steam inlet by 50-80 °C.

In some regimes, a monobloc unit can be started without preheating the steam pipelines, in particular, the regime of starting from the hot state. Moreover, when the thermal insulation is in proper condition, a unit can be started without preheating the re-heater system after a standstill for 1-2 days. A criterion of such regimes is a decrease in the steam temperature by not more than 20-30 °C compared with the temperature at the steam inlet to the turbine. As pre-heating is completed, the flow rate of fuel is corrected so as to ensure the steam-generating capacity of the boiler at which the initial load of the turbo-generator will be around 5% of the rated value. In starting up from a cold or a warm state, the flow rate of the fuel is usually established at the minimal level so as to obtain more easily the required low temperature of live and reheated steam. In starting up from hot state, however, the flow rate of fuel is increased to the upper limit (30% of the nominal value in the circuits with single bypass) to obtain a steam temperature close to the nominal value. Before pushing the turbine rotor, the starting sprays are switched on and the temperature of live steam is established at a desired level. In a once-through boiler, the throttle valve in the line of water recirculation to the deaerator is set so as to obtain a steam pressure before the starting spray nozzles that is 1.5-2.0 MPa higher than the pressure of live steam. In a drum-type boiler, the desired steam temperatures behind the stages of the super-heater are additionally pre-set. In units of higher capacities there is no steam bypass and only the starting sprays to the steam main, which are put in operation before connecting the turbo-generator to the electric network, are used. Before connecting the turbo-generator to the network, when its rotor is accelerated and synchronized, the steam-generator capacity of the boiler and the temperature of live steam are maintained constant. During this period, the temperature of the re-heated steam gradually increases for the same reasons as on connection of the super-heater.

The temperature of the re-heated steam rises especially sharply on connection of the turbo-generator, i.e. when the steam flow rate through the re-heater system is almost doubled. For this reason, it is important to switch on in due time the devices which control the temperature of re-heat. In units with once-through boilers, the start-up and shutdown device is not closed during the period of turbo-generator synchronization, so that all the control valves of the turbine are opened (and heated) due to a drop of pressure in the live steam. In units with drum type boilers, the start-up and shutdown device is partially closed to maintain a constant pressure of live steam, which improves the operating conditions of the boiler drum and control over the steam temperature. As the turbo-generator is connected to the network, the start-up and shutdown device is closed and the unit begins to supply the initial load.

At the third stage of starting (loading) a monobloc unit, its elements are heated from the initial temperature to the final temperature corresponding to its operation under the rated conditions. An attempt to shorten the time of loading may result in the quick heating of the boiler elements and the appearance of high temperature gradients. The highest temperature gradients and highest temperature stresses can appear in massive thick-walled elements, such as the housing and rotor of a turbine, the drum and headers of a boiler, and fittings of the steam mains. In such cases, compressive stresses appear in the heated surface of an element and tensile stresses in its unheated surface. Upon heating an element, temperature stresses diminish to zero and often even change sign. Stresses of an inverse sign can form in elements on the reduction of the steam temperature or shutdown of the unit. With frequently repeated start-up and shutdown regimes, thermal stresses vary cyclically and can cause fatigue cracking of the metal. These stresses determine the allowable heating rates of these elements.

Solid fuel-fired boilers are transferred from starting fuel (gas or fuel oil) to solid fuel, if the load is 15-30% of the nominal value, and the consumption of the firing fuel is gradually reduced. As the boiler begins to supply the specified load, the elements of the starting circuit which are used only at start-up and shutdown are switched off and the voltage supply to the corresponding drive means is discontinued. Some monobloc units, which supply a variable portion of the load curve, are started by an automatic process control system. Such systems can perform automatic control of a given process and carry out discrete operations by means of logic control units. These units an switch on and off the mechanisms of auxiliary system of a boiler, change the position (open or close) of closing valves, switch on and off automatic controller, switch over the controllers from some actuators to others, alter the structural schemes of the controllers, etc. Before performing a particular operation, a logic control unit checks that the operation is allowed. In monobloc units provided with an automatic control system of this type, the boiler operator has following additional functionalities: i) to perform preparatory operations for starting the unit and select the reserve mechanisms which will be switched on automatically; ii) to supervise the functioning of the equipment and perform manual control should a failure occur in the operation of a particular automatic controller; iii) to correct the operating conditions (when needed) by adjusting the set-point devices of automatic controllers; and iv) to check the state of the equipment upon completion of a particular stage of the starting regime and issue commands for a next stage. Thus the automatic control system of a monobloc unit is complex of the technical means of control and the operating personnel which co-operates with these means.

13.16. Friction and Oscillations

Friction and oscillations are closely interrelated, friction is capable of generating oscillations, and oscillations have influence on friction [46]. The discussions are presented here on the concept that the friction force is mainly formed by a load normal to the rubbing surface and by the corresponding normal contact deformation. Important is also the concept that the oscillations in complex elastic systems are interdependent. This means, that the normal and tangential oscillations of a slider cannot arise

irrespective of one another. When one kind of oscillation occurs, the others usually take place too. This interdependence is determined by how close the natural frequencies of the respective oscillations are and by the character of their relation.

Normally directed oscillations generated by friction: If a sliding body or slider has a height that is not much greater than its transverse dimensions, i.e. this slider is other than a rod sliding on its end face, the elastic compliance of the slider in the normal direction is dozen of times lower than the compliance of the contact layer, the latter being a layer formed by the peaks and valleys of micro irregularities of the rubbing surfaces. In this case, the slider is roughly similar to an absolutely rigid body that rests on a number of minute springs which simulate the peaks of surface irregularities. Some of these are in contact and carry the normal load, while others, the smaller ones, are out of contact. Any casual normally directed pulse acting on the slider will cause its free oscillations in the normal direction. These oscillations are non-linear and asymmetrical, because more and more new smaller peaks come into contact as the slider moves downwards, and because the contact rigidity is variable: it rises when the slider moves down and drops when the slider moves up. An increase in the amplitude of these oscillations causes the mean level of the slider over the counter face to rise and thereby the total real contact area to decrease.

As the slider moves along the counter-face, its surface asperities receive micro-pulses from the asperities of the counter-face. The normal components of these micro-pulses continuously generate the oscillations of the slider as a whole in the normal direction. These micro-pulses, chaotic as they are, maintain permanent quasiperiodic sustained oscillations of the slider in the normal direction. These asymmetrical oscillations lift the slider and reduce the friction force. The higher the speed of sliding, the more intensive are the normal components of micro-pulses between the surface asperities, and hence the greater is the amplitude of the oscillations, the higher is the mean level of the slider movement, the smaller is the real contact area, and the weaker is the friction force. That may explain the drooping of the friction force-speed characteristic, i.e. reduction in the friction force with increase in the speed of sliding in the absence of lubrication. Oscillograms of oscillations, which are contact oscillations, show that their amplitude grows linearly with the sliding speed.

Effect of forced oscillations on friction force: To reduce friction, use is made of normally-directed forced oscillations whose frequencies are usually very different from the contact resonant frequency. For this reason a tangible reduction of friction force requires the application of powerful vibrators. The selection of a vibrator is determined by the construction of the tribological joint and by the mass of the vibrating element. Piezo-electric vibrators are convenient to use for small masses, whereas electrodynamic vibrators are used for large masses. The quantitative problem of power rating of the vibrator depending on the slider mass and the rubbing surface properties can also be solved. The power of the vibrator must increase with an increase in its mass. Piezoelectric vibrators are attached symmetrically to the top surface of the slider so that they extend radially from the surface far enough to provide the coincidence of their natural bending oscillation frequencies with the resonance frequency of the contact oscillations and, at the same time, with that of the vibrator. That increases the magnitude of the friction force resonance minimum without changing the frequency. For gradual control, the forced oscillation frequencies should be farther from the resonant frequency.

Normal micro-pulses are proportional to the speed of sliding. With tangential oscillations at work, the normal micro-pulses reach its maximum twice for one period of the tangential oscillations. As there is the natural frequency of the normal oscillations, the frequency of the forced tangential oscillations (which is half that of the normal oscillations), must act as a resonance frequency for inducing the maximum amplitude normal oscillations.

Frictional self-excited oscillations: A non-uniform sliding of solids rubbing against each other under a constant reactive force may involve more or less periodic halts; this uniformity is called frictional self-excited oscillations. A possible explanation for non-uniform sliding is a drooping sliding speed-friction force characteristic. In this case a casual acceleration of sliding reduces the friction force and leads to a further acceleration until the slider has slipped into the position where its external elastic constraints develop a reactive force which causes the slider to slow down. The slow-down increases the friction force and therefore leads to a further slow-down. In some cases the slider slows down to a stop, and then the pull of the elastic drive rises until it dislodges the slider by causing it to skip once again. The non-uniformity occurs when the gradient of the drooping characteristic, i.e. a negative derivative of the friction force with respect of the sliding speed, becomes greater than te coefficient of damping of the drive's oscillating system. The greater this gradient, the more intensive oscillations take place and the higher is the speed below which the slider oscillations are attended with halts. The main practical recommendation for eliminating the non-uniformity of sliding is to increase the damping capability of the slider drive and to lower the gradient of the friction force-sliding speed characteristic or to remove completely the factors which give rise to a drooping characteristic.

13.17. Automatic Control and Management of Noise Systems

Definition of noise systems: The physical structure of a noise system is extremely diverse and complicated. A noise system consists of noise fields of composite wave elements due to physical agents. Noise systems therefore are hard to formalize and deny a uniformly valid straightforward methodology which could be helpful in their analysis and handling. The choice of noise system parameters involves such methods based on experience of handling a noise system by way of system control theory. However, experience obtained in handling one noise system cannot be used in assessing the behaviour of another system, so each noise system has to be described by a separate model. The model evaluator, therefore, has to adopt a multimethod, multidisciplinary approach which naturally, prevents tracing even

general equivalence relations between various noise systems. The introduction of computers brought about a variety of simulation techniques based on the generation of random variables and functions. However, applied to noise system problems, these programs are cumbersome, computer time intensive, and may even fail to warrant optimal solutions. The existent techniques of system control theory – the vector matric approach in the state space and frequency domain techniques – and software package based computational procedures are of limited value for noise system modeling and design [47].

Noise systems operating in the presence of uncertainty or in conflict situations may be handled to advantage by the methods of game theory. These methods are efficient in solving the synthesis problems in noise systems and searching for their optimal system arrangements. Some of the above-mentioned directions pertaining to noise characterization theory for noise system studies have not yet received a due development to become suitable for practical problem solving, mainly owning to the complexity of the respective computational procedures and the need for powerful software packages. Most of these shortcomings can, in all probability, be overcome by establishing super-high efficiency computer aided noise system networks with elements of artificial intelligence. This however, may require a few years of research of many institutions in electronics, instrumentation, computing and information science. The development of appropriate versatile software to support complex design of a noise system may take even longer time. Therefore power design procedures for a noise system capable of optimizing the system arrangements and parameters of noise systems will evolve unless both measurement hardware and simulation and computing software are available.

Wide scale introduction of industrialization, fast transportation systems and modern engineering systems in daily lives of people have led to the development of complex noise systems. Over and above, these noise systems came into existence when they received powerful computers for their control. As a result, a new class of noise systems engineering has appeared in which operating instructions, realized as computer routines, monitor and control continuous subsystems for their automation. Such noise systems are referred to discrete-continuous systems and their design faces certain problems due to the intrinsic complexity and lack of well-established mathematical methods of analysis and synthesis. On the advantage side, computer simulation can be employed even in the absence of analytical methods for noise system design and it is simple for monitoring the time evolution of the noise process concerned. Among the disadvantages, simulation is hard in deriving a model adequate to the actual noise system. Moreover, it cannot generate its own decisions (which can be made with analytical design techniques), and the last but not least, the procedure of noise characterization simulation is rather long as it has to evaluate the effects of a large number of variables, controls, and disturbances. The pursuit for elimination of these shortcomings has led to the development of analytical techniques, facilitated by the recent advancements in computer hardware and software along with internet of things (IoT).

The flexible automated lines, for example in industry and transportation systems are a very broad class of noise systems. They offer improved output at minimum costs due to an organizational approach, which is directed towards mechanized systems, which lack in optimization and control of noise from these systems. This approach takes account of all the ingredients of production cycle including raw material supply, manufacturing, and transportation of finished items for further processing and assembly, or toward warehouses. This implies that these automated lines are multistep noise systems with a given subsystem arrangements. The intrinsic constraints for such noise systems due to composite wave elements (viz., noise fields) are energy and materials balances in the form of inventories in stock, production resources and storage capacity for raw materials and display capacity for ready products.

The mathematical models for this type of noise systems take the form of multivariate algebraic equations and inequalities. Optimal parameter search under large number of noise system constraints is achieved by high capability software programs with linear and non-linear variables. To simplify computational procedures, both the direct and decomposition techniques assume that the matrix of constraints is rather sparse, i.e., contains many zeros and linear terms. In the direct methods based on linear programming and the simplex method, the main computational procedure is the manipulation of the inverse of the base case. To compute the inverse matrix it is decomposed into a product of elementary matrices each containing maximum possible number of zeros. The inverse matrix can also be represented as a product of two triangular matrices, which significantly reduces the amount of computations. If the algorithms of storing the matrix are such that the columns are stored with the use of special markers introduced into the code, then what is stored is a reduced-size matrix of elements alien to the initial matrix.

Decomposition methods represent a large system as a number of individual subsystems. Decomposition can exploit the techniques for linear and nonlinear compensation of interconnections, and the sequential elimination of low-impact controls and disturbances. These methods are normally realized in the form of structural transformations. The computational procedure of large problem decomposition is built around column generation. Columns are generated where necessary, but none of them is stored so that the main storage requirements are considerably reduced. The decomposition method is an iterative procedure comprising a sequential solution of the noise systems problem and its subsystems involved. Once the initial feasible basis of the noise system problem is generated, one can derive the dual variables to determine the objective function of its subsystems. Solving the subsystems in turn generates new columns for the noise system, which give rise to a new basis and dual variables and the computational procedure is done with recycled iteration.

The above mentioned process is a simplex algorithm which converges to an optimal solution in a finite number of iterations. For cycling of iteration, additional manipulations are necessary to ensure convergence of the iterative process. With an appropriate

modification of the problem formulation, the decomposition principle can be used to solve problems of large and complex noise systems. New control languages and data processing codes introduced for solving large scale linear programming and mixed integer programming problems are well adapted for computer aided modeling and control of noise systems. Synthesis of flexible automated noise flow fields and noise control systems should allow for time dependent functions governed by the probabilistic behaviour of input constraints and their associated service times. These noise systems are multilevel arrangements with an arbitrary structure, and their modeling may be achieved with advantage on the lines of queueing theory. Because of the complicated structure of the processes in noise systems the modeller has to allow for a multitude of various factors, the complete list of which is not always attainable. A much higher efficiency than in deriving analytical dependences can be achieved with asymptotic techniques. These exploit the fact that physical parameters of the system can be treated as small variables. In the context of modeling a noise system, asymptotic techniques perform in much the same manner as exact methods: both yield dependencies which form a base for exhaustive search algorithms of the parameters essential for the system development.

Noise management systems would require the need for introducing priority servicing strategies. As a rule, priority is assigned on the basis of service time. Another approach is to assign priority in accordance with the value of the service-time to waiting-cost ratio, so that the highest priority is assigned to the arrival of service with the smallest ratio. Numerous other methods may be invoked for priority assignment. However, in noise systems, a priority strategy does not always result in analytical relationships for modeling, forcing the choice of parameters to be performed by complex methods of mathematical modeling.

Modeling studies in queueing theory have received wide application areas in noise systems comprising of communications, computer networks, transportation systems, equipment maintenance scheduling, resource management etc. It should be noted that many areas of the current research in queueing theory resulted because of the advancements of computer information systems. This expanded the range of solvable problems, and gave rise to new methods of decision making helpful in deriving arrangements of noise systems for their modeling, simulation and control.

There exist a class of noise systems operating under uncertainty and conflict conditions. Each conflict situation taken from the practical experience is normally extremely complicated and its analysis is hampered by a multitude of secondary factors. Modellers would require analyzing such noise systems with simplified game models. The salient feature of these models is that none of the players is informed in advance on the decision made by the others, i.e., the player has to choose his strategy under uncertainty. At times the uncertainty of the outcome may be due to a priori unknown external disturbances. The game theory models involved in the solution of noise system synthesis include finite and infinite zero-sum games, cooperative games and differential games. Finite non-cooperative games are used to study discrete noise systems, or continuous noise systems which can be reduced to the discrete counterparts. In synthesis problems, utility functions are formulated in terms of payoff matrices. In n-person games, each player has its own utility function.

In infinite zero-sum games, players tend to reach antagonistic objectives and the set of feasible solutions is infinite. This feature of the games makes them a suitable tool for the synthesis of continuous noise systems. Cooperative games are involved when the noise system problem allows pre-play coalitions concerning players' strategies to be set up, and the payoff to be distributed among the players on a pre-arranged basis. Differential games are, essentially, multistage games with continuous time. The transition from state to state is then described by a set of differential equations. The usual assumption is that two players together control the motion of a particle in n-dimensional space, and controls for this motion are chosen by means of a vector or scalar performance criterion. Game theory models are fairly unwieldy and have come into use for system synthesis after information and computing system science became available. For most practical problems the analyst has to invoke heuristic approaches and interactive optimization systems.

Computer simulation: In complicated noise systems, dynamic automatic control can be described by high-order differential or difference equations. In order to arrive at the precision and quality of the performance specification for a noise system modeled, the modeller has to adjust parameters many times so the procedure may run into long simulation times. The modules of a computer simulation program in the noise characterization field can be compiled by means of a vector matrix description of dynamics in state space. Matrix representation in frequency domain can be very useful for algorithms computing transfer functions, frequency characteristics, transient processes, and studying the dynamic accuracy of continuous and discrete automated noise control systems. This approach to analysis and synthesis of complicated automatic noise control systems may relieve the modeller from routine operations. The engineer or modeller engaged in computer simulation normally has to analyze many sets of different alternatives for the noise system to choose one which most fully meets the performance specifications. At the final stage of analyses, procedures of search theory and decision making theory may be invoked. The chosen alternative is decided by two factors: i) by the probabilities of the outcomes that would take place; and ii) by the priorities used in comparison of alternative outcomes. Most widely used techniques rely on decision making with priority subject to a given objective function. These techniques are supported by the respective algorithms of quasi-optimal or optimal decisions. The complexity of automatic noise control systems decreases the modeller's capability of treating all the information concerning priority in a non-formal manner. Therefore, new decision making techniques are developed which are based on judgmental assessments. These assessments are generated as the modeller subjects various system alternatives to his judgement with account of numerous relevant factors such as precision, quality, cost, reliability, weight and size, and most importantly noise pollution control factors and the

information on the interconnection of these factors. In addition, the technique takes into account the judgemental assessments of experts in the noise characterization and control field. Since these assessments are performed by several experts, the designer ranks them and evaluates correlated judgements, and the level of significance for the performed expertize. If the deviation of the judgemental assessments is insignificant, then the correct decision may be evaluated with a very high significance coefficient. On the other hand, deviation of judgements may be used to advantage in a collective assessment. Collective brainstorming sessions may bring about interesting results, without a total comparison with the individual priorities derived. In all the situations considered, decision analysis offers advantages over analytical techniques because it allows formalized account of numerous subjective factors essential for the problem on hand. Some decision making techniques rely on multi-critical methods of optimal control. In some situations, the multi-criteria problem of optimal control reduces to one with a scalar criterion. The further development of decision making is seen on the way of implementing the methods of finite games using vector criteria.

To date decision making has reached a status when it can handle complicated problems of noise system design. However to ensure correct decisions to be made more versatile methodological procedures are needed, that are built around multifactor priorities and collective actions of experts, as well as the related computer simulation software. The above mentioned methods of noise system analysis enable the modeller not only to obtain optimal arrangements, but also to determine their numerous characteristics of a noise system. Solution to the noise system problem encounters considerable difficulties because of inadequate information on the capabilities of the existent noise systems either because of a small size of the respective system's family or because of total lack of the respective analogs. Practical prognosis includes several methods such as the analogs technique, trend extrapolation, the development model, construction trees of objectives, and statistical estimation. As a rule these methods are backed up by individual or collective expert assessments. Certain disadvantages have been noted for these methods. More specifically: The analogs technique is difficult to derive identical characteristics. The trend extrapolation cannot ensure a correct choice of the characteristics being forecast not to say about a large amount of statistical data required for its realization.

The best forecasting technique is the statistical method because it ensures a high reliability of prognoses at a rather scarce initial data input. However its use incorporates a reduction in the space of characteristics. The significance of a particular characteristic for the noise system's potential evaluation depends on its contribution to the accuracy of the noise system's class identification. Normally this contribution is measured in terms of losses, specified by a loss matrix. The statistical approach also requires a model which separates all evaluated noise systems into feasible and infeasible categories. Then the potential of the noise system may be assessed from the probability of the parameter vector in each of the classes concerned. However, often the scarcity of initial data prevents this operation from being implemented. Therefore, the characteristics are combined in groups. The number of combinations is to be determined from modeling. These combinations are used to estimate the accuracy of the classification and the loss matrix is set-up. If the expected losses computed with such matrices are minimal, then the system may be classed as feasible one. After this, it is assumed again that each combination is a signature and they are classified. The process continues until the accuracy of classification ceases to alter.

For automatic noise control systems, consideration should be given to the non-linearity involved and the random nature of the signals. The behaviour of such noise systems is described by stochastic differential equations. When these equations run into high dimensionality their solution encounters significant difficulties. The analyst has to invoke various linearization techniques for nonlinear functions, including statistical techniques. It should not be overlooked, however, that these models are valid so long as the assumptions fundamental to the statistical linearization principle hold true. This limitation can be removed by invoking the path integral technique which expands the range of nonlinear problems handled. Basically the path integral technique consists in representing the solution of nonlinear stochastic equation as a sum over all feasible paths of the dynamic noise system. The set of these paths can be finite or enumerable for discrete systems, and have the power of continuum for continuous noise systems.

The state of art in noise system theory is such that the necessary prerequisites are existent for solving problems of the noise management of multilevel economic systems. These problems include integration of noise generation activities, scheduling, transportation planning, optimal facilities location and layout and project management. For their solution, one may exploit systems analysis and certain methods and their state of art algorithms. Most of these are heuristic techniques which yield only approximate solutions. The quality of solutions improves when the problems are solved systematically in an iterative manner so that final decision is made by the modeller. Further development of noise system theory can be motivated by the cumulative number of problems which can give rise to a variety of research efforts tailored to suit numerous applications.

Automatic control in noise systems: The main direction of current technological development would be to control the process variables in the noise system and arranged the system as per desired objectives. The field is however materially penalized by lack of research in the development of rigorous mathematical methods which would allow the modeller to select the most efficient arrangement and parameters for constitutive modules of the noise system. In noise systems, which are made for automatic controlling of either complicated or simple objects, the streams of data and controls may be of essentially diverse physical nature, which significantly hampers the formulation of mathematical models and consequently the development of its design and modeling methodology. The most complicated group of noise systems are those involving as one of the control links a human operator or a group of people coordinating, matching, or

predicting control actions.

Actual noise systems consist of a wealth of individual interdependent subsystems of various levels, of hierarchy. The behaviour of these subsystems varies under the impact of the environment and control instructions issued for the system to accomplish its objectives. A multilevel arrangement of the noise system leads to the need for a sequential stage-by-stage coordination of activities within the noise system. The efficiency of such coordination depends on the choice of local controls at each stage such that it ensures an optimal performance of the noise system. The process of sequential coordination with iterative procedure can be used for improving the global performance of a noise system.

The number of levels used in devising the hierarchy of a noise system is conditioned by the adopted mathematical model. In an elementary description the arrangement may be treated as consisting of a number of interrelated subsystems of the noise system. The hierarchical arrangements can pave the way for a general approach to their buildup. All the arrangements have a vertical system with the control actions formed at the upper level flowing downwards. Three basic types of multilevel noise systems, most frequently occurring in practical situations are as follows: i) with vertical flow of control actions; ii) with controls applied via feedforward and feedback loops; and iii) with the use of matched control actions in feedback loops. The modellers need to breakdown such noise systems into simpler components with the use of decomposition techniques. The decomposition is known to boil down to the synthesis of a noise coordination system modification device. Accordingly, the condition central to a decomposition is that the noise system is coordinated. Once the noise system is decomposed into separate subsystems, the synthesis problem is solved for each of them and the optimal arrangement and parameters derived that satisfy certain deterministic local criteria. This is done in such a way that the recombination of so derived subsystems into a single noise system may ensure extremity of the global deterministic performance measure formed from the adopted local criteria by some elementary operations. Then the parameters of subsystems are varied to arrive at the given stochastic performance specification. It should be recognized that this approach is valid only for noise systems described by differential and difference equations reducible to linear equations. Its applicability to nonlinear noise systems requires additional validation based on the methods of operations research. Two step optimization – first with the deterministic and then with the stochastic performance measures – considerably simplifies the estimation of effectiveness and modeling of noise systems.

In the absence of random external disturbances and equipment failures, all multilevel noise systems can be synthesized with a deterministic performance criterion allowing only two outcomes (if the noise system accomplished the objective or not) and subject to a number of constraints imposed on the time of accomplishing all the stages of the operation. However, actual noise systems operate in random environments, under the conditions of randomly varying parameters. For recurrent operating cycle the effectiveness of design alternatives can be assessed using stochastic performance measures.

Noise control system and its characterization: The increasing complexity of noise control systems has led to the call for their refinement of the formal models used in their design. Such a model describes in a formal manner how the process being controlled interacts with the environment. In the classical context of automatic control, including a controlled process, a decision making block, and feedforward and feedback loops, a control action is generated from information on the controlled process and its state, the data on the state arriving through the feedback loop. If the controlled process is a continuous noise system, it is described by laws specifying its functions as a rule in the form of integro-differential equations in which the state of the controlled process is a continuous function of time. From the practical viewpoint, however, controlled processes in the form of discrete systems are characterized by: i) a finite set of events that transfer the noise system from one state into another (including the transitions that occur as a result of control actions); ii) a finite set of conditions defining noise system's states; and iii) a finite set of controls. Such a noise system normally denies a description by equations including time functions. A decision on a control is made upon assessing the current state of the object. Such an estimate is a function of the parameters defining the arrangement properties of the noise system rather than its numerical characteristics, as is the case with continuous noise systems. The noise system may contain resources that cannot be shared – a typical situation in practice – then a request for this resource may not be honored immediately and the process will have to wait for this resource a certain time. The management of a system thus reduces to allowing certain processes to hold resources and taking a resource away from a process, so that the process is blocked. The quality of management is then decided by the number of processes in the wait condition and the number of resources being held by blocked processes, i.e., waiting processes and resources held by them are undesirable.

Noise management systems: A variety of methods may be involved in building a model for a multilevel noise system with hierarchical control, called a noise management system. These are simulation techniques, queueing theory, various optimization techniques, game theory etc. The methods are combined in order to achieve more or less comprehensive models. Such multifaceted approach hampers establishing the interconnection between subsystem descriptions and, when in an industrial context, the erection of a model for enterprise noise management. These problems owe their existence in the large part to the absence of a clear cut global comprehension of the system being controlled, specifically of the noise system structure. Noise management system modeling presents the analyst with particular problems.

Assume that the noise system to be modeled is a transportation network. What should be meant by the element in such a noise system? Since the task of transportation is to provide the shipment of goods, it would be logical to assume that this element is a complex of three types of facilities, namely, loading means, carrier vehicles, and unloading means. Such a complex

aggregated for a shipment disintegrates when the process of shipment is over and its components are available for the formation of new complexes which in turn disintegrate. Thus the mentioned noise system will be conceived as a multitude of elements being integrated and disintegrated in a random fashion. Consequently, it is difficult to reveal regularity and to define the noise system structure. However, for the noise system to perform its main function, consistent with the objective for which it has been devised, it is immaterial what the specific complexes will be, i.e., which facilities will be integrated in them. This implies that there must be a noise system structure that is sufficiently stable and adequately defines noise system components.

Every noise system can be described either in the lines of transforming the input excitations into the output quantities, or by way of a process aimed at achieving a certain objective or performing a certain function. The first approach is referred to as the cause-effect process. Essentially, it decomposes the noise system on hand into simple elements interconnected by cause-effect ties according to the input-output principle. Each element is to perform a certain, logically completed, specific function, the union of which, upon interconnecting the elements under the set principle, is to achieve the global function of the noise system. Exciting the input of such an element by a standard waveform, one may evaluate how the element functions by recording its output and describe this function by a model. All these activities identify the physical element having input and output ports with the function it performs. By identifying an element or module with the function it performs, elements are represented by different structures – physical and functional. This identification is valid until cause-effect, or input-output, connections can be traced out in the noise system and this is decomposable into elements in accordance with this principle. Unlike engineering, physical and other types of system, noise management systems are never closed, therefore their evolution or purposeful development is an integral factor of their existence. Owing to this fact, the control of the development of such noise systems is the most important type of control as it defines not only the improvement of the management system but also the quality of its performance and its very existence.

13.18. Research Areas in Noise Characterization Theory

Human noise behaviour: a) Developing a slide rule for noise measurement; b) Acoustic/Noise Filters, built environment analysis, human comfort & health through selective acoustic/noise filter configurations; c) Physiology and epidemiology of human noise behaviour.

Integrated control of built environment parameters for comfort and wellness: Providing acceptable Indoor Environmental Quality is a critical function of built environment system through characterization and measurement (odors, dust, pollen, light, humidity, heat, sound, electricity, airflow, air temperature, fire etc.). With integrated environmental control, these environmental parameters are to be characterized and checked through sensation and perception of occupants for comfort and wellness and controlled through various environmental monitoring sensors. The effect of these human behavioural parameters is to be characterized and measured on a logarithmic noise scale.

Distributed/Noise generation characterization & modelling; health effects (+/-); comfort/discomfort indices: a) Aero-acoustics/wind-induced natural ventilation; b) Pedestrian comfort/discomfort/noise indices with wind-induced corner peak pressure on tall buildings; c) Solar & wind energy conversion acoustics on PV/T collectors; d) Hybrid wind and solar PV systems (off Grid) for serving remote city areas; e) Wind energy conversion acoustics/electromagnetic interference due to wind mills; f) Heat transfer, thermal compression oscillations, thermo-acoustics.

Energy conversion materials characterization and noise insulation: a) Novel solar energy conversion materials, their properties & characterization for selective configurations of solar photovoltaic, thermoelectric and solar thermal collectors; b) Integrated insulation materials properties; c) Heat transfer, energy, entropy & noise generation modelling.

Energy conversion materials and noise generation modeling: a) Solar energy conversion by biological & photosynthetic materials; b) Hydrogen production through solar energy conversion; c) Biomass/Waste-to-Energy Conversion; d) Energy, exergy, thermodynamic, entropy & noise generation modelling.

Energy conversion (materials, devices) and noise: a) Investigating solar energy acoustics by using/devising geothermal energy to supply power of a mega power size plasma torch for plasma gasification of waste; b) Investigation on their control system using wireless power transfer; c) Energy, combustion, exergy, thermodynamic, entropy & noise generation modelling.

Industrial energy conversion (materials, devices) and noise management: a) Industrial energy conversion, waste/biomass combustion/incineration and cogeneration for electric power and HVAC comfort requirements; b) Energy conversion in engines and noise reduction; c) Noise reduction in transportation systems; d) Noise systems modeling and control.

Instrumentation: a) Sensors & transducers for measurement, monitoring and control of noises (light, sound, heat, fluid, electricity, fire and the sun); b) Electromagnetic, thermoelectric and piezoelectric transducers for noise measurement.

Graduate teaching courses: a) Building Insulation; b) Sustainable Eco-Cities; c) Human Environmental Health; d) Noise Systems Engineering; e) Occupational Cellular Physiology; f) Solar Energy Acoustics; g) Measurement Systems and Units.

14. Conclusions

A study on acoustic signal processing and noise characterization theory via energy conversion in a PV solar wall device with ventilation through a room is performed. The noise interference and characterization as per speed of a composite wave is presented. The sources of noise waves (sun, light, sound, heat, electricity, fluid and fire) are described depending on their speed of noise interference. Noise measurement equations and their units are coined. The power systems are classified as per

source signals of solar power, electric power, light power, sound power, heat power, fluid power and fire power. The noise filters for filtering noise from power systems are defined with examples.

Energy conversion is investigated in a PV solar wall by means of experiments and numerical model. The outdoor experiments were conducted on a PV solar wall installed on a wooden frame connected to a room with passive and active ventilation. Further, outdoor experiments were established under available operating conditions. The experiments conducted on a PV solar wall have taken into consideration the effect of heat capacity, thermal time constant, and thermal storage losses. The experimental results along with results of simulation model for noise filtering for a PV device are presented.

The electrical and thermal characteristics were established for a PV solar wall. The sensible heat storage capacities of a PV solar wall installed on a wooden frame were higher in comparison to a window glass or a stand-alone PV module. The electrical and thermal characteristics were function of both outdoor air temperature and thermal storage losses. Finally, production of heat and electricity was a nonlinear function of volume of electrical and thermal resistances developed across a PV solar wall similar to a loudspeaker. The active ventilation was deemed to be essential in order to avoid over-heating of a PV solar wall and to maintain adequate outdoor ventilation in a room under mild climate conditions.

Some noise unit examples for an air duct exposed to solar radiation are tabulated. A phenomenon of photovoltaic amplification for a pair of photovoltaic modules connected to a potentiometer is enlightened. The time plots of power function were used to support and devise noise measurement expressions and noise characterization in a power system as per speed of a wave.

A simple theory is outlined in this paper which gives an indication of the way in which noise due to physical agents are attenuated by a complex system or building elements. In practice due to complex elements, the noise transmission due to composite waves comes both via the common partition panel and the balance from the supporting structure. Theory and measurement applicable to panels can be extended to complex building systems, since it is the objective to devise the lightest and simplest panels with the highest values of insulation. It is the net effect of the panel partition with its associated structure which decides the insulation obtained in practice and which is difficult to predict.

Detailed discussions of noise characterization theory in various disciplines are presented through energy conversion model and experiments with emphasis on cities, acoustics, electrical, control, noise fields, energy conversion in engines, noise pollution & control, fire alarm systems, noise instrumentation, operation of power systems, friction & oscillations along with some discussions on automatic control & management of noise systems. The paper has also introduced theory and concept of solar energy acoustics. Description of sensors & transducers for a human brain with example of human comfort & health are also presented. Research areas in noise characterization theory along with its graduate teaching subjects are briefly outlined.

Appendix has presented some basic definitions to understand the noise characterization theory.

Acknowledgment

Part of the work was conducted by the author at Department of Building, Civil and Environmental Engineering, Concordia University, Montréal, Québec, Canada.

References

[1] H. Dehra, **"**A Paradigm for Characterization and Checking of a Human Noise Behavior", WASET Proceedings of 19th International Conference on Psychological and Brain Sciences, pp. 658-666, May 11-12, 2017, Montréal, Canada.

[2] H. Dehra, **"**Electrical and Thermal Characteristics of a Photovoltaic Solar Wall with Passive and Active Ventilation through a Room", WASET Proceedings of 19th International Conference on Sustainable and Renewable Energy Engineering, pp. 584-592, May 11-12, 2017, Montréal, Canada.

[3] H. Dehra, "A unified theory for stresses and oscillations", *Proc. Canadian Acoustical Association (CAA) Montréal 2007 Conf*, Concordia University, Montréal, Québec, Canada , pp. 132-133, 2007.

[4] H. Dehra, "Power transfer and inductance in a star connected 3-phase RC circuit amplifier", *Proc. AIChE 2008 Spring Meeting,* New Orleans, LA, USA, session 96a, 2008.

[5] H. Dehra, "The noise scales and their units", *Proc. Canadian Acoustical Association (CAA) Vancouver 2008 Conf*, Vancouver, B.C., Canada, pp. 78-79, 2008.

[6] H. Dehra, "A benchmark solution for interference of noise waves", *Proc. AIChE 2009 Spring Meeting*, Tampa, FL, USA, session 67c, 2009.

[7] H. Dehra, "A guide for signal processing of sensors and transducers", *Proc. AIChE 2009 Spring Meeting*, Tampa, FL, USA, session 6b, 2009.

[8] H. Dehra, "Solar energy absorbers", chapter 6 in *Solar Collectors and Panels, Theory and Applications*, edited by Reccab Manyala, InTech Publication, pp. 111-134, 2010.

[9] H. Dehra, "Acoustic filters", chapter 5 in *Ventilation: Types, Standards and Problems* edited by Vincent A. Romano and Allison S. Duval, Nova Publishers, pp. 135-154, 2012.

[10] H. Dehra, "A theory of acoustics in solar energy", *Natural Resources*, pp. 116-120, 4 (1A)**,** 2013.

[11] H. Dehra, "A novel theory of psychoacoustics on noise sources, noise measurements and noise filters"*, Proc. NoiseCon16*, Providence, Rhode Island, pp. 933-942, 13-15 June, 2016, Publisher: Institute of Noise Control Engineering, USA, 2016.

[12] H. Dehra, "On Sources and Measurement Units of Noise", Proc. International Conference on Innovation, Management and Industrial Engineering (IMIE 2016), Kurume, Fukuoka, Japan, 05-07 August 2016, 219-227, ISSN: 2412-0170, 2016.

[13] H. Dehra, "A Slide Rule for Noise Measurement", 10th International Conference on Sustainable Energy Technologies (SET 2011), Istanbul, Turkey, September 4-7, 2011, 5 p.

[14] V. Del Toro, *Electrical Engineering Fundamentals*, 2nd ed., Prentice-Hall of India, New Delhi, India, 1986.

[15] Robert H. Randall, *An Introduction to Acoustics*, Dover Publications, USA, 1951.

[16] Ingo R. Titze, *Principles of Voice Production*, Prentice Hall, Englewood Cliffs, NJ, USA, 1994.

[17] J. C. Lucero, "Dynamics of the vocal fold oscillation", *TEMA Tend. Mat. Apl. Comput.*, 6, No. 1, pp. 11-20, 2005.

[18] J.L Threlkeld, *Thermal Environmental Engineering*, Englewood Cliffs, NJ, USA, 1962.

[19] H. Dehra, "*A heat transmission model for a telephone line",* Proc. of 21st CANCAM, Department of Mechanical and Industrial Engineering, Ryerson University, Toronto, Ontario, Canada, June 3-7, 2007, pp. 356-357, 2007.

[20] H. Michael Newman. *Direct Digital Control of Building Systems,* John Wiley and Sons, New York, USA, 1994.

[21] H. Dehra, **"**Photovoltaic Solar Wall: 2-D Numerical Modeling and Experimental Testing under Fan induced Hybrid Ventilation", proceedings of the IEEE International Conference on Energy Efficient Technologies for Sustainability, April 7-8, 2016, St. Xavier's Catholic College of

Engineering, Nagercoil, Kanyakumari District, Tamilnadu, India, IEEE Xplore, ISBN: 978-1-4673-9925-8, Page(s): 668-676.
[22] H. Dehra, "A Multi-Parametric PV Solar Wall Device", in proceedings of IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017), 978-1-5386-0814-2/17/$31.00 ©2017 IEEE, Chennai, India on Sep 21-22, 2017, pp. 392-401.
[23] H. Dehra, "Characterization of Noise in Power Systems", proceedings of IEEE International Conference on Power Energy, Environment & Intelligent Control (PEEIC2018), 978-1-5386-2341-1/18/$31.00 ©2018 IEEE, Greater Noida, India on April 13-14, 2018, pp. 321-330.
[24] H. Dehra, "A numerical and experimental study for generation of electric and thermal power with photovoltaic modules embedded in building façade," submitted/un-published Ph.D. dissertation, Dept. Building, Civil and Environmental Engineering, Concordia University, Montréal, Québec, Canada. August 2004.
[25] H. Dehra, "Experiments on photovoltaic modules embedded in building façade," AIChE Spring 2010, San Antonio, TX, USA, March 21-25, 2010.
[26] H. Dehra, "The effect of heat and thermal storage capacities of photovoltaic duct wall on co-generation of electric and thermal Power," AIChE 2007 Spring Meeting, Houston, Texas, USA, April 22-26, 2007, session 36a.
[27] C. Muresan, C. Ménézo, R. Bennacer, R. Vaillon, "Numerical simulation of a vertical solar collector integrated in a building frame: radiation and turbulent natural convection coupling," Heat Transf Eng, pp. 29-42, 27, 2006.
[28] J. H. Kim, J. T. Kim, "A simulation study of air-type building-integrated photovoltaic-thermal system," Energy Procedia, pp. 1016-1024, 30, 2012.
[29] Y. Lin, C. Chiang, C. Lai, "Energy efficiency and ventilation performance of ventilated BIPV walls," Eng Appl Comput Fluid Mech, pp. 479-486, 5, 2011.
[30] C. J. Ho, A. O. Tanuwijava, C. Lai, "Thermal and electrical performance of a BIPV integrated with a micro encapsulated phase change material layer," Energy Build, pp. 331-338, 50, 2012.
[31] H. Dehra, "The effect of heat and thermal storage capacities of photovoltaic duct wall on co-generation of electric and thermal Power," AIChE 2007 Spring Meeting, Houston, Texas, USA, April 22-26, 2007, session 36a, 9 pages.
[32] H. Dehra, "The Entropy Matrix Generated Exergy Model for a Photovoltaic Heat Exchanger under Critical Operating Conditions", International Journal of Exergy, Vol. 5, Issue 2, pp. 132-149 (2008).
[33] H. Dehra, "A Two Dimensional Thermal Network Model for a Photovoltaic Solar Wall", *Solar Energy*, 83 (11) 1933-1942 (2009).
[34] H. Dehra, "An investigation on energy performance assessment of a photovoltaic solar wall under buoyancy-induced and fan-assisted ventilation system", Applied Energy, pp. 55-74, 191, 1 April 2017.
[35] H. Dehra, "A Combined Solar Photovoltaic Distributed Energy Source Appliance", Natural Resources, June 2011, Issue 2, pp. 75-86.
[36] H. Dehra, "A Mathematical Model of a Solar Air Thermosyphon integrated with Building Envelope", International Journal of Thermal Sciences, Vol. 102, April 2016, pp. 210-227.
[37] ASHRAE Handbook, *Fundamentals Volume* (1997), American Society for Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA, USA, 1997.
[38] ANSI/ASHRAE Standard 55-2004. *Thermal environmental conditions for human occupancy*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA, USA, 2004.
[39] Jerrold H. Krenz, *Energy Conversion and Utilization, 2nd Edition,* Allyn and Bacon, Inc,, Boston, USA, 1984.
[40] Larry Gay, *The Complete Book of Insulating,* The Stephen Greene Press, Vermont, USA, 1980.
[41] V. Lukanin, *Internal Combustion Engines,* Mir Publishers, Moscow, 1990.
[42] H.J. Purkis, *Building physics: Acoustics*, Oxford, New York: Pergamon, 1966.
[43] H J Purkis, "Instruments for noise measurement", Journal of Scientific Instruments, Volume 41, Number 5, pp.284-289, 1964.
[44] M.K. Poltev, *Occupational Health and Safety in Manufacturing Industries*, Mir Publishers, Moscow, 1985.
[45] M. I. Reznikov and Yu. M. Lipov, *Steam Boilers of Thermal Power Stations*, Mir Publishers, Moscow, 1989.
[46] I.V. Kragelsky, *Friction, Wear, Lubrication – Tribology Handbook*, Vol. 2, Mir Publishers, Moscw, 1986.
[47] A. A. Voronov, *Management and Control in Large Systems*, Mir Publishers, Moscow, 1986.

Appendix

Definitions

Energy Area: The definitions of noise sources are characterized by energy area stored in a wave with its speed and difference due to power intensities of two waves due to interference.

Noise of Sol: The difference of power intensities between two solar power systems causes noise of sol (S). The power storage on a unit area per unit time defines amplitude of a solar energy wave. The storage of solar power is defined by a solar energy wave pack of unit cross sectional area and of length s, the velocity of light.

Noise of Therm: The difference of power intensities between two heat power systems causes noise of therm. The power storage on a unit area per unit time defines amplitude of a heat wave.

The storage of heat power is defined by heat energy wave pack of unit cross sectional area and of length s, the velocity of light.

Noise of Photons: The difference of power intensities between two lighting power systems causes noise of photons. The power storage on a unit area per unit time defines amplitude of a light beam.

The storage of light beam is defined by light beam packet of unit cross sectional area and of length s, the velocity of light.

Noise of Electrons: The difference of power intensities between two electrical power systems causes noise of electrons. The power storage on a unit area per unit time defines amplitude of an electricity wave.

The storage of electrical power is defined by an electricity wave pack of unit cross sectional area and of length s, the velocity of light.

Noise of Scattering: The difference of power intensities between two fluid power systems causes noise of scattering. The power storage on a unit area per unit time defines amplitude of a fluid wave.

The storage of fluid power is defined by fluid energy wave pack of unit cross sectional area and of length s, the velocity of fluid.

Noise of Scattering and Lightning: The difference of power intensities between two fire power systems causes noise of scattering and lightning. The power storage on a unit area per unit time defines amplitude of a fire flash.

The storage of fire power of light is defined by fire pack of unit cross sectional area and of length s, the velocity of light. The storage of fire power of fluid is defined by fire pack of unit cross sectional area and of length s, the velocity of fluid.

Noise of Elasticity: The difference of power intensities between two sound power systems causes noise of elasticity. The power storage on a unit area per unit time defines amplitude of a sound wave. The storage of sound power is defined by sound energy wave pack of unit cross sectional area and of length s, the velocity of sound.

Nomenclature

T_p	Temperature of PV module
T_b	Temperature of insulating panel
T_a	Temperature of air
S	Solar Intensity
E_p	Electric power
V	Air velocity
T_o	Ambient air temperature

T_s	Room air temperature
$H_{pv\text{-}T}$	Heat capacity
ΔT_v	Temperature difference (y-ordinate)
ΔT_H	Temperature difference (x-ordinate)
ρ_n	Density
K_d	Thermal conductivity
H_d	Film coefficients
T	Thermal time constant
Q_v	Thermal storage capacity (y-ordinate)
Q_H	Thermal storage capacity (x-ordinate)
C_n	Specific heat
d_n	Thickness
$d_n\rho_nC_n$	Energy stored/m^2/K
I_{sc}	Short circuit current
V_{oc}	Open circuit voltage

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 140-151 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

SDN-based Network Control Method for Distributed Storage Systems

Luis Guillen[*,1], Satoru Izumi[1], Toru Abe[1,2], Hiroaki Muraoka[3,4], Takuo Suganuma[1,2]

[1]*Graduate School of Information Sciences, Tohoku University, 980-8577, Japan*

[2]*Cyberscience Center, Tohoku University, 980-8577, Japan*

[3]*Tohoku University, 980-8577, Japan*

[4]*Tohoku Institute of Technology, 982-8588, Japan*

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 12 September, 2018
Online: 28 September, 2018

Keywords:
SDN
Distributed Storage Systems
Network Management

ABSTRACT

With the increasing need for effective storage management due to ever-growing content-generation over the Internet, Distributed Storage Systems (DSS) has arisen as a valuable tool. Although DSS has considerably improved in the past years, it still leverages legacy techniques in its networking. To cope with the demanding requirements, Software Defined Networking (SDN) has revolutionized the way we manage networks and can significantly help in improving DSS network management. In this paper, we propose an SDN-based network control method that is capable of handling DSS network management and improving its performance. This paper presents the design, implementation, and evaluation using an emulated environment of a typical Data Center Network (DCN) deployment. The experiment results show that by applying the proposed method, DSS can increase the performance and service resilience compared to existing solutions.

1. Introduction

This paper is an extension of a previous work originally presented at the 2017 International Conference on Network and Service Management (CNSM2017) [1]. User-generated content is growing exponentially. From 30 Zettabytes (ZB) of content generated in 2017, it is foreseen to reach 160 ZB by 2025 [2]. Although end-users create most of this content, with the widespread use of the Internet of Things (IoT) and the advances in cloud technologies, content generation will also increase at the core of the network. It is also worth noting that in 2017 approximately 40% of the content was stored in enterprise storages [2], but due to the paradigm shift from expensive and large data-centers to cloud-based virtualized infrastructures, it is also projected to increase to 60% by 2025. The massive scaling and flexibility required in those infrastructures will demand more efficient ways to handle the additional traffic.

At the outlook of such demanding requirements, Distributed Storage Systems (DSS) became more popular, since they provide highly reliable services by networking nodes to provide enhanced storage [3]. Over the years, DSS has progressively achieved better performance by improving propagation and recovery methods, from simple replication to more advanced techniques [4,5]. However, little attention has been paid to improvements at the network level, as they still rely on legacy techniques.

To cope with the increasing need for efficient storage managed by DSS, in this paper, we propose a network control method that is capable of handling the generated traffic by specific DSS tasks. The proposed method is based on Software Defined Networking (SDN) [6], which separates the control plane from the data plane and will allow a more flexible programmable network. The contribution of this paper is to show the potential of applying this paradigm to improve DSS performance. Furthermore, we describe the inherent problems of DSS when using legacy techniques in Section 2, and the minimum requirements that DSSs demand from the network perspective, namely aggregated bandwidth and practical use of resource. Based on those requirements, we designed the proposed method as described in Section 3, whose main strength is its simplicity. More concretely, we define a solution to handle the bandwidth aggregation called *on-demand inverse multiplexing*, and we detail the solution for the practical use of resources called *multipath hybrid load balancing*.

[*] Corresponding Author: Luis Guillen, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan, +810222175080, Email: lguillen@ci.cc.tohoku.ac.jp

To test the feasibility of the proposed method, in Section 4 we evaluated the implementation of typical DSS scenarios, the results showed that our method outperforms traditional ones, and is capable of delivering the required features by DSS.

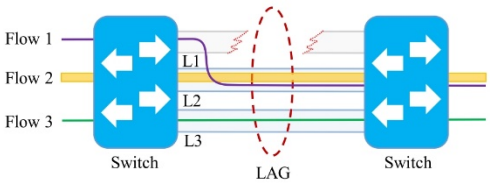


Figure 1: Programmability problem with LAG

2. Related Work

2.1. Problem Description with Legacy Networking Techniques

To the best of our knowledge, solutions that combine features of both technologies (DSS and SDN) have not been fully explored yet, however, in this section we present related work on individual features.

Initially, we should mention that there are two minimum requirements DSS will need from the networking perspective, namely: aggregated bandwidth and practical use of resources. In this sub-section, we describe the specific problems with legacy techniques when applied to DSS regarding these two aspects.

In the case of bandwidth aggregation, since DDS nodes might be allocated in different storage servers, a bottleneck is created at the server gateway when various clients try to access it at the same time, due to the limited bandwidth of a single link. Kaneko et al. [7] tried to overcome this problem by using Link Aggregation (LAG) [8], which allows bandwidth aggregation by grouping a limited number of physical links as a single logical bundle. LAG offers communication resilience by redirecting the incoming traffic to another active link in case of failure. However, there is no control on the selection process, e.g., in the simple LAG deployment depicted in Figure 1, three links (L1, L2, and L3) are grouped in a single bundle, if L1 fails then the protocol redirects the traffic to another link, but the traffic may well be sent through the most congested link (L2) instead of the L3 which has less traffic. Moreover, for a link to be part of a bundle, all of them need to have the same configuration, and it is limited to a hard-coded number of links, which significantly limits the flexibility of the system.

In the second case, namely the practical use of resources, DSSs are typically deployed on Data Center Networks (DCNs) using common topologies, such as a three-layer non-blocking fully populated network (FPN) or three-layer fat-tree network (FTN) [9]. In these environments, network traffic management is usually leveraged to techniques such as Equal Cost Multipath (ECMP) [10]. However, ECMP is not efficient regarding resource usage, and since DSS clients access several storages concurrently, congestion mostly occurs at some segments of the network. Additionally, since it is limited to legacy protocols such as Spanning Tree Protocol (STP), it will prune redundant links, which can be used for creating alternative paths. For example, in the topology shown in Figure 2, if two flows go from H1 and H2 to H3, some routes might be preferred (e.g., the red dashed path), despite the availability of other paths. Moreover, as shown in Figure 3, a bottleneck is generated at the segment nearest to the end-device, for instance, if H5 have to reply flows from requests send by H1-H4, the link connected to the nearest switch will be highly congested (i.e., red dashed path). Apart from the number of

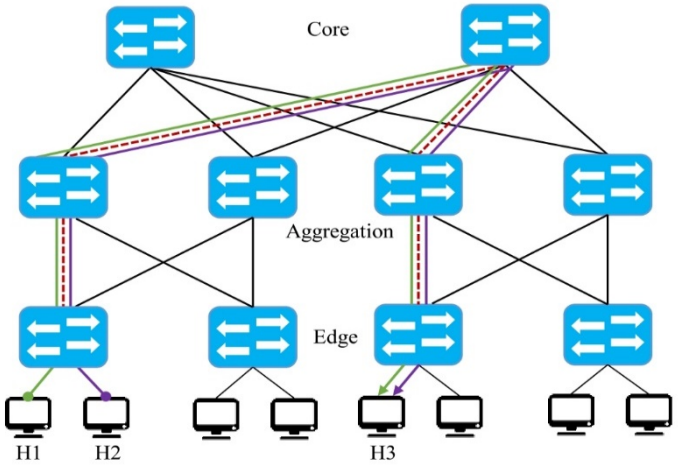


Figure 2: Traffic congestion issue in DCNs

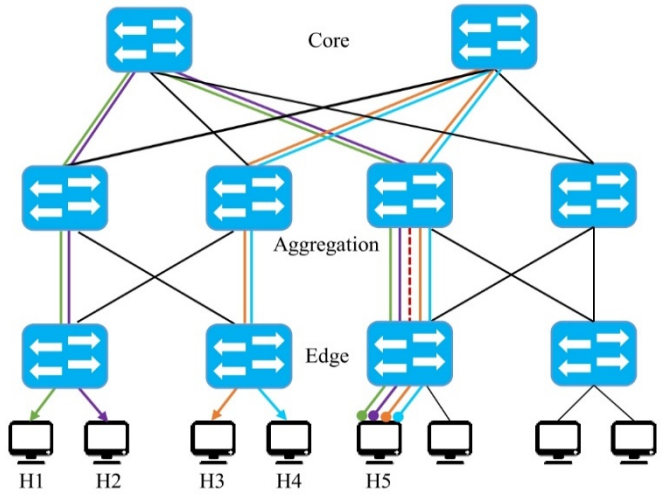


Figure 3: The last-mile bottleneck issue in DCNs

flows going through a single link, that particular segment of the topology the connection speeds are not as fast as they are in the core, we call this problem “last-mile bottleneck.”

To sum up, legacy techniques cannot provide DSS with neither aggregated bandwidth or efficient use of resources due to the last-mile bottleneck issue, and limitations with end-to-end routing programmability with techniques such as ECMP.

2.2. Related work on SDN-based Multipath Load-balancing

The primary task towards a resource-efficient method is to balance the load among the available resources, and using multiple paths is a practical way to relieve the network traffic. Thus, in this section, we present load balancing solutions that focus on multi-path solutions using SDN.

Initially, it is worth noting that load balancing is a topic that has been extensively explored for decades, but SDN-based solutions are relatively more recent. SDN-based load balancing can be

categorized depending on their architecture in Centralized and Distributed [11], we focus on the centralized architectures wherein the management is performed in the data plane by a central controller that has an overview of the entire network. In this context, a pioneer work was *Hedera*, a dynamic flow scheduling for DCN capable of balancing the load among the available routes. The main contribution of the authors is the introduction of two placement heuristics to allocates flows, and the estimation of flow demands that uses OpenFlow (OF) for routing control. However, the authors did not consider last-mile bottlenecks nor the load on the server side.

OLiMPS [13] is a real implementation of Multipath TCP (MPTCP) [14] in an intercontinental OF-network that achieved high throughput. The drawback of this work is that, as other solutions [15, 16], they rely on MPTCP, which is an experimental protocol capable of handling resource pooling using multiple paths. However, the inherent problem is that it needs modifications in the end-point kernel. To avoid this restriction, Banfi et al. [17] presented MPSDN, a multipath packet forwarding solution for aggregated bandwidth, and load balancing that achieves similar results to MPTCP with the added value of not requiring end-point modifications. Their main contribution is the idea of including a threshold called Maximum Delay Imbalance (MDI), used to place flows in paths. However, even if they do not require end-point modification, they need to modify Open Virtual Switch (OVS).

DiffFlow [18] used a selection mechanism to handle flows in a DCN, so that short flows will be handled by ECMP and long ones using a process called Random Packet Spraying (RPS). However, by partially relying on ECMP they inherit the same issues concerning the use of resources.

Li and Pan [19] proposed a dynamic routing algorithm with load balancing for Fat-tree topology in OpenFlow based DCN. Their flow distribution strategy provided multiple alternative paths from a pair of end nodes and placed the flow to the one with the highest bandwidth. However, the hop-by-hop recursive calculation overlooked the available bandwidth of the entire network. A similar approach was presented by Izumi et al. [20], who proposed a dynamic multipath routing to enhance network performance, they introduce an index based on the risk and the use parallel data transmission and distribute the traffic using multiple paths. Dinh et al. [21] also presented a dynamic multipath routing, capable of selecting *k*-paths to distribute the traffic based on the load of the links; however, they only handled the initial assignment, and the selection of the number of paths is unclear.

Finally, Tang et al. [9] present an OF-based scheduling scheme that dynamically balances the network load in data centers. Their approach aimed to maximize the throughput by designing a heuristic based on the available resources in the network. A significant contribution of their work is that they present a theoretical model for load-balancing and specific metrics for network utilization and load imbalance. The drawback, however, is that they heavily rely on the use of a pre-calculated table (ToR Switch-to-ToR Switch Path Table S2SPT) for path selection, which dramatically limits the dynamicity in case of real implementation, as the computation time will increase if the number of links is relatively high. Moreover, they are still subject of the last-mile bottleneck problem due to the limitation in the DCN topology.

2.3. Target Issues

From the related work presented in the preceding sub-sections, and considering DSS requirements, we summarize our target issues as follows.

(P1) Last-mile bottleneck: Links have lower connection speeds at layers closer to the end-device in a DCN topology causes a network bottleneck. Therefore, DSS performance will be limited when performing parallel tasks.

(P2) Limited use of multipath routing: Despite the available redundant links in the topology, paths from end-to-end nodes are usually mapped as single-paths via some preferred routes, this provokes unnecessary congestion in specific segments of the network.

(P3) Load Imbalance: Due to the limited use of the available resources DSS neither the network nor the servers are used in a balanced manner. The existing solutions focus on either one of those aspects, which leads to overlooking the importance of both variables for an effective control method.

3. SDN-based Network Control Method for Distributed Storage Systems

3.1 Motivating example

We describe the proposed SDN-based network control method to solve the problems described above. However, before describing the proposed approach, let us consider a simple scenario of a DSS process.

Suppose we have a DSS that uses network codes [3] to recover failed nodes. In this scenario, the information of a node (D) is fragmented into p=4 pieces (p_1, p_2, p_3, p_4) stored in different nodes from the set of nodes (N), such that any piece can be reconstructed from any k=3 pieces in N, but it needs the four pieces to recover D. If a node fails, the DSS needs to conduct two processes, namely, regenerate the piece and resume the primary data transmission.

The recovery process is described in Figure 4. Initially, *D* collects information from p_1, p_2, p_3, and p_4, but imagine p_4 fails; then the system needs to identify and locate the other pieces (p_1, p_2, and p_3), establish the corresponding recovery links (marked with dashed lines) and transfer the recovery data from the surviving pieces to a new node (p_4'). Once the transmission of the regenerated piece is over, the newly created piece (p_4') will need to continue the primary data transmission, to do so, the system needs to identify the location of the new piece, establish the new link (blue line), and resume.

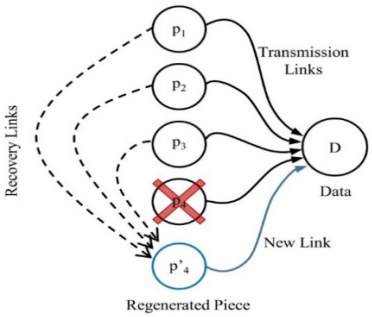


Figure 4: Sample process in DSS

As observed, even in this small example, various processes occur from the networking point of view. Needless to say that the efficiency of the DSS will depend on how fast a node can be recovered and how efficiently uses the resources.

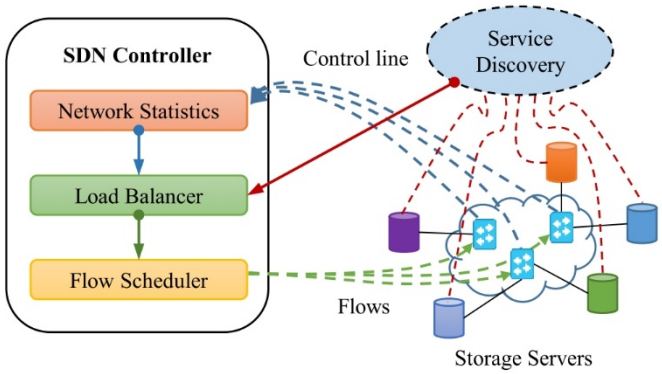


Figure 5: Overview of the Proposed Method

3.2 Overview of the Proposal

To solve the target issues, we describe the proposed SDN-based Network Control Method for DSS. The overall scheme is depicted in Figure 5. As observed, the architecture consists of a centralized SDN controller that is comprised by three internal modules: *Network Statistics*, *Load Balancer*, and *Flow Scheduler*; additionally, an external module (*Service Discovery*) interacts with the controller and the storage servers connected in the underlying network. The role of each of these modules is described as follows:

- *Service Discovery* - this module is in charge of tracking the status of all the storage servers. This information includes the physical load of each of the servers (e.g., memory, CPU, number of processes being served), and the specific DSS configuration (i.e., the number of pieces and their location). The interaction with the SDN Controller is direct, and the status is sent when requested.

- *Network Statistics* - periodically collects network statistics of the entire network, i.e., network topology changes, transmitted and received packets per port. This polling process will allow having an overview of the whole network infrastructure before calculating the appropriate paths. Although the polling happens periodically, it can also be triggered directly by request. We use this information to calculate parameters such as the available bandwidth (1)

$$b_{i,j} = \frac{(Rx+Tx)_{current} - (Rx+Tx)_{previous}}{\Delta t} \quad (1)$$

- *Load Balancer* - this is the central module in charge of calculating the paths based on the information collected by the *Service Discovery* and *Network Statistics* modules. The primary goal of this module is to distribute the traffic among the available paths based on not only the state of the network but also the load of the storages containing the required pieces. The overall process is subdivided into three phases, namely Discovery, Load Balancing, and Adaption (as depicted in Figure 6). In brief, once the request is sent from a source, in the Discovery phase, this module calculates the available paths from source to destination with a process we call *on-demand inverse multiplexing* (described in Section 3.3). Then, based on the current status of both network and storage servers, in the Load-balancing phase, a process we call *hybrid Multipath Load-balancing* assigns the paths that best suit the request. Finally, in the Adaption phase, once the transmission has started the status is updated periodically in case better routes become available or if any change occurred in the topology. The main load-balancing procedure is described in Algorithm 1, where a control loop checks if there has been any change in the topology within a fixed period. As observed, the function *adjustWeights()* in line 8, requests an update from the *Network Statistics* module, and in line 9 the function *balanceTraffic()* is in charge of performing the load-balancing.

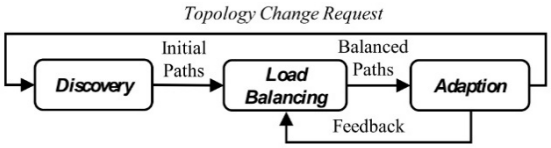


Figure 6: Load Balancing Steps

- *Flow Scheduler* - once the most suitable paths have been selected, this module writes the flow rules into the specific network devices that take part in the transmission process.

Algorithm 1: Topology Monitor

```
1:  function monitorTopology ();
2:  startTimer(t);
5:  do
6:      if hasTimeElapsed() then
7:          if hasTopologyChanged() then
8:              adjustWeights(),
9:              balanceTraffic(client, Servers, NW),
10:         else restartTimer() then
11:         end
12:     else addTimeSpan();
13:     end
14: While true;
```

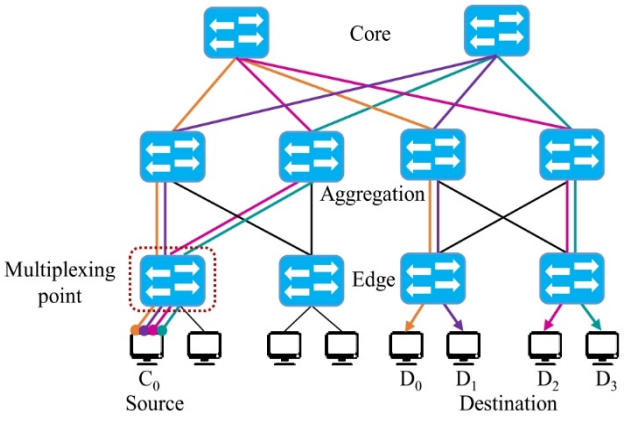


Figure 7: On-demand Inverse Multiplexing

3.3 On-demand Inverse Multiplexing

As mentioned in the previous subsection, the initial step towards an effective network control is the path discovery and traffic distribution. In this section, we describe a process we called *On-demand Inverse Multiplexing*. Initially, it is worth noting that *inverse multiplexing* was explored in the past [22] and mainly used for bandwidth aggregation. It refers to the simple process of sending traffic over multiple paths. In essence, our particular contribution is the idea of distributing the traffic among a specific number of paths (k), which depend on the requirements in a particular DSS configuration. However, in a typical DCN deployment only single links are assigned from device-to-device; moreover, connections at the last layer in the topology are usually low-speed connections, which create a phenomenon we call a *last-mile bottleneck*. To solve this problem, we propose to augment the topology by using parallel connections from the layer closest to the end-devices (edge layer) to the next hop (aggregation layer). This topology augmentation is central for our overall proposal and would involve the overprovision of links according to the particular configuration of the DSS, i.e., setup *n* links if the system split data in a maximum of *n* pieces. Although it is not a common suggestion, organizations still prefer to use various cheap links instead of expensive limited services, such as MPLS or single fat-line [23] due to the cost constraints. Note that, by performing this augmentation, the power consumption of the involved network devices will increase proportionally to the number of parallel connections, therefore, it is necessary to think of a strategy that is capable of reaching a tradeoff between the performance and the power consumptions, however, such a strategy is out of the scope of this paper.

To illustrate the implementation of the on-demand inverse multiplexing process, let us consider the scenario depicted in Figure 7, in this simple deployment, a Fat-tree fully-connected topology comprises three layers (Edge, Aggregation, and Core). Furthermore, assume the particular DSS uses four pieces out of 5 to regenerate as the maximum level of division. In this case, the number of links that the network needs to provision from the edge to the aggregation layer is four. Of course, the augmentation need not be among each device, but only in those particular segments wherein the requirements are higher and can be configured statically or calculated on-demand based on the configuration in the DSS gathered by the *Service Discovery* module.

Once we have the infrastructure prepared, the remaining process works as follows. Consider again the scenario depicted in Figure 7, where we emulate the data recovery which will be allocated in a storage server (C_0), the pieces needed for recovery are located in different servers (D_0, D_1, D_2, D_3). Therefore, in this simple example, the number of parallel links needed are k=4 paths. When the request arrives at the multiplexing point (the closest Top of Rack TOR switch), the controller calculates the appropriate number of paths (*k*) based on the information at the *Service Discovery* module and the current state of the network captured by the *Network Statistics* module.

To solve the initial path discovery, we calculate the Maximum Disjoint Paths (MDP) by a process described in Algorithm 2. This process, which is the initial step of this work and was introduced in [1], calculates the MDPs by an adapted version of the Suurballe's [24] algorithm which calculates the path candidates. Note that we use a multigraph $G = (V, E)$, where V is the set of network devices (switches), and E is the set of links. Since it is a multigraph, more than a single link can connect two nodes in the set V. We assume that the bandwidth (b_i) for each edge is symmetric, which means that both the uplink and the downlink are the same. Moreover, each edge has a cost $\mu_{u,v}$ as shown in (2), where B is the set of all individual bandwidths b_i. This value is updated periodically by the *Network Statistics* module.

$$\mu_{u,v} = \frac{b_i}{max(B)} \tag{2}$$

Algorithm 2: Path selection algorithm to find the kmax disjoint paths from source to destination

```
1:  function selectKPaths (s, t, k, G);
    Input  : s, t, k, G(the Vthe ,E)
    Output: Set of k path candidates P
2:  P ← Ø;
3:  currentPath ← Ø;
4:  nPath ← 1;
5:  do
6:      if nPath > 1 then
7:          adjustWeights (P[nPath-2]),
8:      end
9:      currentPath ← getDijkstraShortestPath(s,t)
10:     if currentPath ≠ Ø then
11:         nPath++;
12:         P.add(currentPath);
13:     end
14: While currentPath ≠ Ø and nPath ≤ k
15: return P
```

3.4 Hybrid Multipath Load-balancing

This part of the paper was partially presented in a previous work [25]. Usually, load-balancing techniques focus on either server or network load. In the proposed method, we consider both variables to balance the traffic, and that is why we call it *hybrid*. Moreover, since the topology was augmented with parallel links, it is necessary to take into account multiple paths to balance the traffic. The main procedure is described in Algorithm 3. Initially, from a pool of servers *S*, the *lookupServers* function (line 2) searches for the server candidates which can provide the required service (e.g., in the case of a regeneration process in DSS, it locates the nodes containing the pieces necessary to regenerate the current piece), which are then ordered based on the load of the node. In line 8, we search *k* alternative paths from the source *c* to each of the server candidates *s* that can provide the service. In line 9, the function *assignPaths* calculates the minimum combination of server and path cost. It is worth noting that, based on initial experimentation, the difference in overall cost from all the paths should be less than 25% otherwise the transmission will be delayed in the paths that have costs with the higher difference. This process will guaranty that the distribution is homogeneous in both, the network and servers. Finally, once the destination servers have been identified and the paths selected, the function *writeFlows* in line 11 will send the balanced paths to the Flow Scheduler module, which sends the instructions to the network devices involved in the path.

Algorithm 3: Hybrid Load Balancing Algorithm to calculate the best paths

```
1:  function balanceTraffic (c, S, G);
    Input : Source, set of Servers, G(V,E)
2:  serverCandidates ← lookupServers(c,S);
5:  if isNotEmpty(serverCandidates) then
6:      foreach s in serverCandidates do
7:        k← s.lenght
8:        Candidates ← selectKPaths(c,s,k,G)
9:        BalancedPaths←assignPaths(c,Candidates),
10:     end
11:     writeFlows (BalancedPaths);
14: end
```

To illustrate the selection process, consider the contrived topology depicted in Figure 8. In this particular example, the features of the system are as follows:

- C_0 is a newly created node that needs to restore the data, to do so, it requires four pieces (p_1, p_2, p_3, and p_4)
- From the edge layer, there are two connections to the aggregation layer
- All links in the topology have weight=1
- None of the servers have any activity yet, and therefore the load is 0%.
- The storage servers that contain the replicas, among others, C_0 are S_1, S_2, S_3, S_4
- Each request consumes 25% of the server resources, and 100% of the link capacity, and therefore only one of the request can be served per path

When the recovery process starts, the *lookupServers* function will identify that S_1-S_4 have the pieces and their location, and all the candidates will be generated in pairs {(C_0, S_1), (C_0, S_2), (C_0, S_3), (C_0, S_4)} and since there are four pieces (p_1, p_2, p_3, and p_4) the function *selectKPaths* will search for four paths between each pair. Then the selection process in function *assignPaths* will start assigning the appropriate paths. For example, p_1 in C_0 to S_1, via the path: $a \rightarrow b \rightarrow e \rightarrow h \rightarrow g$, adjust the weights and the load percentages and continue the process recursively until all the pieces have a service provider; the final result will look as shown in Figure 9, in which all the servers have a perfect balance, and the network traffic is evenly distributed among the available paths. Of course, this is not the unique process, neither are the request homogeneous, therefore, in case another process starts, i.e., from sources *X*, *Y*, or *Z* the values will vary in the service providers and the network load.

Note that this is just the initial assignment, but in case there is a change in the topology (e.g., a link disconnection, or a network device fails), the whole process needs to adapt and restart the process. In the proposed method, the adaption can happen proactively and reactively. In the first case, as shown in Algorithm 1 in Section 3.2, a timer determines how frequent the system needs to control the changes. Although fail-tolerance is not the focus of this paper, the value needs to have a trade-off between performance and

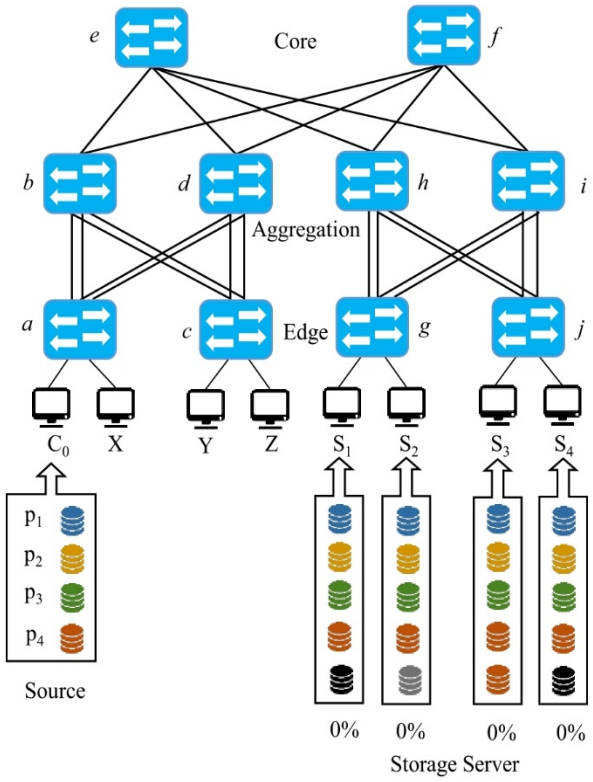


Figure 8: Sample hybrid load balancing (Initial state)

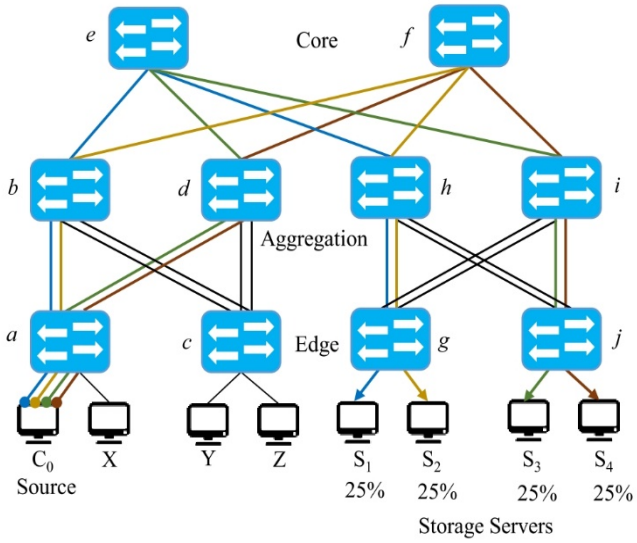


Figure 9: Sample hybrid load balancing (Final state)

service resilience, since having a small refresh time in the order of milliseconds would offer better service resilience at the cost of performance due to the constant polling. In case of reactive adaption, which is an event-based procedure triggered by any change in the topology during the transmission, the system needs to recalculate all the paths in the affected segments.

3.5 Implementation

Based on the overall scheme presented in the previous sub-sections, we implemented the proposed approach using a commonly

used SDN controller OpenDaylight[1] Belirium-SR3 (ODL), OF 1.3 as the communication protocol, and OVS as the back-end deployment.

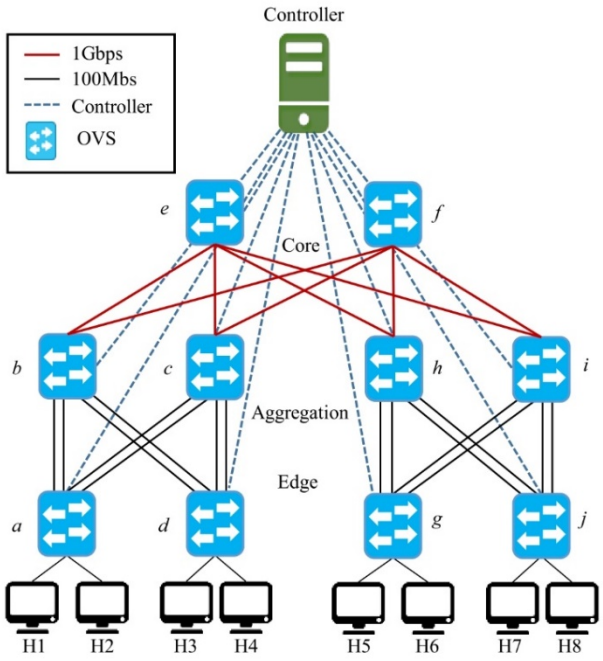


Figure 10: Testbed Experiments

Each of the modules was implemented as part of the controller as follows. The *Service Discovery* module provides a global record of servers' status so that every time a new request arrives at the controller it verifies and updates the status of all the servers involved, and when the transmission is over the status is once again updated. This module could keep track of various parameters, but for simplicity in the current implementation, we only use the following attributes.

ServerID, MaxLoad, Ports[ID, CurrentRequest, Location]

MaxLoad is the maximum number of requests that the server can handle, which we then use to calculate the server load based on the individual requests, and finally the parameter *Ports* is an array of ports that represent a different service, note that for each of the ports we store the current state and where they can be located.

The *Network Statistics* module, collects the network variables every 10 seconds in the current implementation, although the polling time can be configured directly in the controller if the time is too short; correspondingly, the number of control messages increases. The initial topology discovery is conducted by *L2Switch*, which is a feature available in ODL that handles, among others, the ARP handling, host tracking, and so forth. However, once the initial connectivity is ensured, all path decisions will be made by the *Load Balancer* module.

Finally, the *Load Balancer* is integrated a separate feature and is triggered every time a new request arrives at the controller. Once the paths are selected, the *Flow writer* module will send the *flow_mod* messages to the devices and set up the proper flow rules with an expiration time equal to the refreshing time, so that in case the flows are not in use in a cycle they will be deleted from the table.

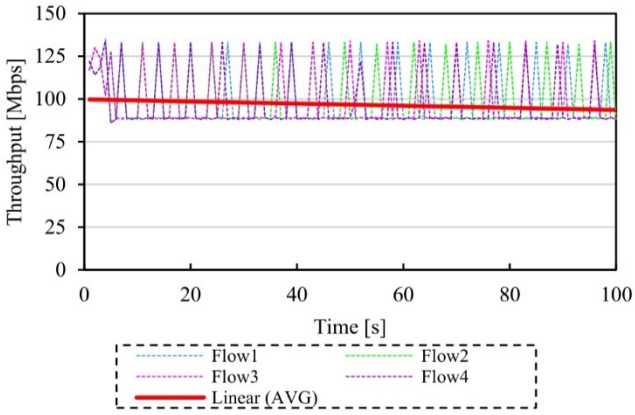
4. Evaluation

4.1. Overview

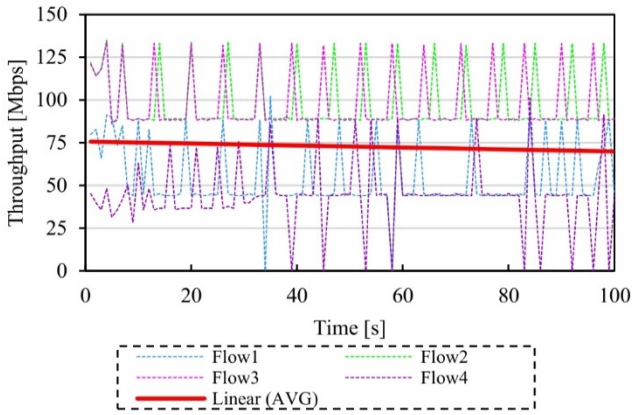
To evaluate the proposed approach, we used an emulated environment created using Mininet[2] v.2.2.2, ODL Beryllium-SR3 as the controller, *iperf* [3] to create the network traffic, and Wireshark to analyze the results. Initially, we explored the correctness and behavior in controlled conditions in which the only traffic is the stream being tested. Then, we performed measurements on environments where other transmissions are happening concurrently and compare the performance with existing solutions.

4.2. Transmission without Background Traffic

In this section, we describe a set of experiments that demonstrate the correctness and effectiveness in conditions where there is no background noise, which means that no other transmissions are happening at the same time. Figure 10 depicts the evaluation



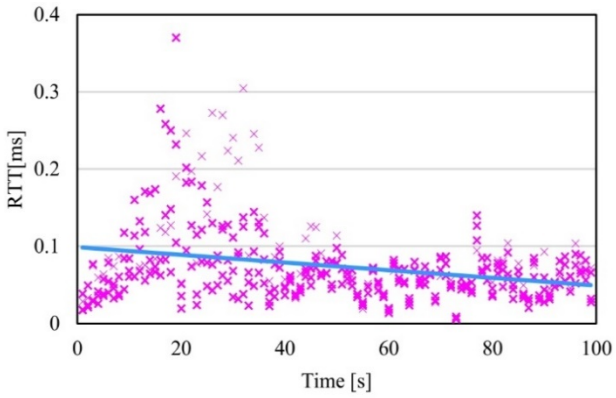
(a) Average throughput one-to-one test



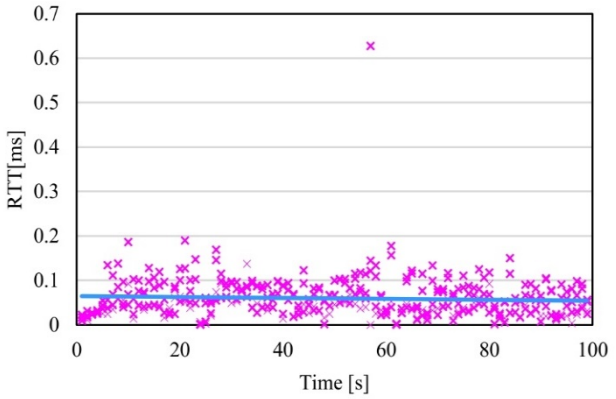
(b) Average throughput one-to-many test

Figure 11: Results average throughput

[1] https://www.opendaylight.org/
[2] http://mininet.org/
[3] https://iperf.fr/



(a) Average RTT one-to-one test



(b) Average RTT one-to-many test

Figure 12: Results average RTT

setup of the testbed used in the evaluation. As observed, a DCN is deployed in a fully connected Fat-Tree-based topology with eight servers, each containing four pieces represented as TCP ports. Each of the links connecting the *edge* to the *aggregation* layer was configured with 100Mbps, whereas the links connecting the *aggregation* to the *core* layer were set up at 1Gbps. Note that the topology was augmented by four parallel links which connect the devices from the edge layer as the number of pieces is also four, each of them will be named as the assigned letter and a subscript index based on the connection port, e.g., the link a_1 means OVS *a* port 1. Also, the SDN application and the emulated network were hosted in the same Virtual Machine using Ubuntu 16.04 LTS, with two 2.60GHz CPUs and 4 GB of memory.

- One-to-One Transmission Test

The goal of this first experiment is to test the correctness of the proposal when the traffic distribution is continuous from a single source to a single destination. In DSS, this case applies when peer-to-peer storages share the same information as simple replication. To conduct this experiment, we used *iperf* to send a continuous TCP stream to four different TCP ports from H1 to H8 for 100s. Note that just for testing purposes no delay was set up for any of the links in mininet.

Initially, the path discovery assigned the following paths {[$a_1 \rightarrow b_1 \rightarrow e \rightarrow h_3 \rightarrow j_3$], [$a_2 \rightarrow b_2 \rightarrow f \rightarrow i_3 \rightarrow j_3$], [$a_3 \rightarrow c_1 \rightarrow e \rightarrow h_4 \rightarrow j_4$], [$a_4 \rightarrow c_2 \rightarrow f \rightarrow i_4 \rightarrow j_4$]}, which is the ideal distribution this particular case. Figure 11a shows the throughput achieved by each of the flows, note the linear trend (in red) is within the best theoretical threshold for each link. Then, we also measured the average RTT of the entire stream, the results are shown in Figure 12a which is relatively high at the beginning of the stream as there is a single destination, but it is still within the boundaries of the standard parameters.

Table 1 shows a summary of the obtained results. From all the flows a total of 4.5Gbytes were transmitted in the 100s period, and the average of throughput was 96.6 Mbps with an average overall RTT of 0.07ms as expected since most of the lines were expedited.

- One-to-many Transmission Test

In this experiment, the continuous stream was sent from a single source to multiple destinations. In DSS, this case applies when a recovering a failed node or when reconstructing the data. To conduct this experiment, we sent an *iperf* request from H1 to a different TCP port in H5-H8 for 100s. A summary of the obtained results is described in Table 2. The discovered paths were {[$a_1 \rightarrow b_1 \rightarrow e \rightarrow h_1 \rightarrow g_1$] for H1, [$a_3 \rightarrow c_1 \rightarrow f \rightarrow i_1 \rightarrow g_3$] for H2, [$a_4 \rightarrow c_2 \rightarrow e \rightarrow h_3 \rightarrow j_1$] for H7, and [$a_1 \rightarrow b_1 \rightarrow f \rightarrow i_4 \rightarrow j_4$] for H8}, note that there is an overlap in the first segment of the paths which affected the overall throughput, as shown in Figure 11b, note the linear trend (in red) achieved 25% less than in the previous case. After tracing the costs of the paths, we found out that there were two paths with the same cost and the algorithm assigned the one first discovered, which shared common segments. Nonetheless, the loss in throughput and the amount of data transferred (see Table 2) was compensated with a more homogeneous overall RTT, as seen in Figure 12b.

Table 1: Results Experiment 1 (one-to-one)

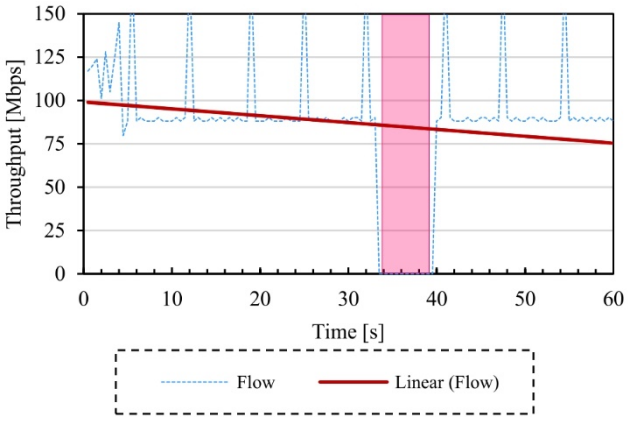
Flow#	Data [Gb]	AVG TP [Mbps]	AVG RTT[ms]
Flow 1	1.13	96.6	0.09
Flow 2	1.12	96.7	0.05
Flow 3	1.13	96.6	0.07
Flow 4	1.13	96.7	0.07

Table 2: Results Experiment 2 (one-to-many)

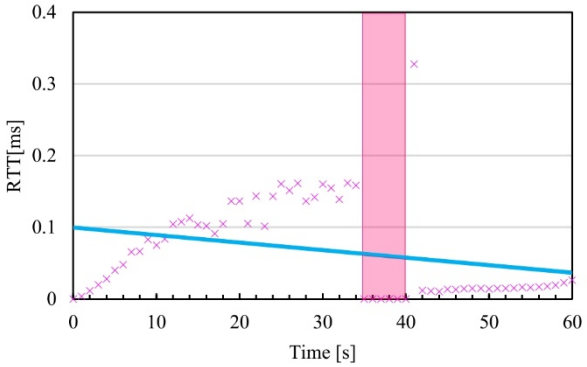
Flow#	Data [Gb]	AVG TP [Mbps]	AVG RTT[ms]
Flow 1	0.65	53.0	0.04
Flow 2	1.13	96.5	0.08
Flow 3	1.12	96.5	0.08
Flow 4	0.52	42.8	0.04

- Path Adaption Test

In this experiment, we test the resilience of the service by sending a continuous *iperf* request to a single server for 60*s* from H1 to H8, and after a random period, one of the links was shut down, Table 3 shows the results obtained. As observed, in the 60s a total of 624Mbytes were transmitted, with an average throughput of 85.6Mbps. Moreover, Figure 13 depicts the transmission trend, which was relatively stable. The initial path was calculated as follows $a_1 \rightarrow b \rightarrow e \rightarrow h_4 \rightarrow j_2$, then the link $h_4 \rightarrow j_2$ failed and recovered later on to a new path $a_4 \rightarrow c \rightarrow f \rightarrow i_4 \rightarrow j_4$, the outage time (shown in the red shaded part in the figure) was approximately 5s. Although the time could be reduced if only part of the path was modified, in practice it is more efficient to recalculate the whole path than look for the part that is affected as the time to add a flow is much lower



(a) Average throughput single source and destination



(b) Average RTT single source and destination

Figure 13: Results path adaption test

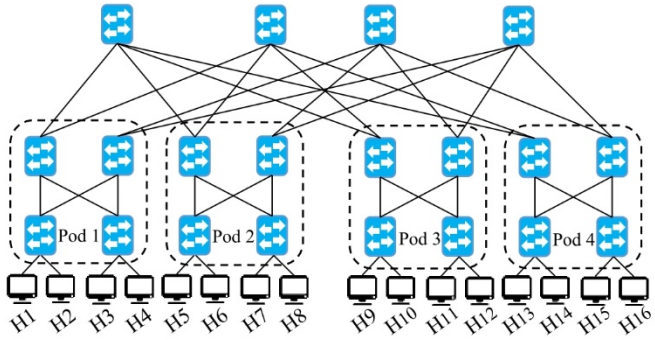


Figure 14: Evaluation testbed 4-ary Fat-tree topology

than when it is updated [26]. Moreover, since the other flows will not have activity anymore, they will be removed in the next refresh cycle.

4.3. Transmission with Background Traffic

In this section, we compare the approach with existing related work. Figure 14 shows the evaluation environment, which consists of a k-ary Fat-tree topology with K=4 pods. Eight servers (Pod 3 and Pod 4) will provide four storages represented as TCP ports. Without losing generality, for testing purposes, all of the links were configured to 100Mbps with a *max_queue_size* of 1000. Note that the topology was not augmented for this test, as the traditional techniques and the existing solutions could not handle the increased amount of redundancy, as the percentage of loss and drop packets was high. However, in the comparison, we also include the results in the case where the topology was augmented.

Table 3: Results Experiment 3 (one-to-many)

Flow#	Data [Mbytes]	AVG Throughput [Mbps]	AVG RTT [ms]
Flow 1	624	85.7	0.06

To benchmark the approach, we compared the results with the following:

- Single path; the deployment was entirely handled by the L2Switch project in ODL, which uses the STP protocol and Dijkstra's algorithm to calculate the shortest path for full connectivity.
- ECMP; An implementation of ECMP using a modified version of a publicly available code[4], that uses OF *group rules* to switch traffic when there are multiple paths.
- MPSDN [17]; the code is also publicly available[5], both the controller (Ryu[6]) and the mininet were hosted in the same virtual machine using Ubuntu 14.04 LTS (as the modified version of OVS was initially designed for a particular version of the kernel). We selected this work, as they proved that for certain conditions their performance equals the one of MPTCP, and thus, we implicitly benchmark our approach against this experimental protocol as well.
- Our approach; The deployment for our approach without modifying the topology, which means that no extra links were setup between the *edge* and *aggregation* layer in the topology. Moreover, the initial paths remained the same for the entire transmission, with no adaption.
- Extended version of our approach; For the extended version of the proposed approach, we modified the original Three-layer non-blocking FTN by adding an extra parallel link from the *edge* to the *aggregation* layer, and the paths were adapted after every 10s (refresh cycle) based on the information given by the *Network Statistics* module.

In the benchmark test, each host in Pod 3 and Pod 4 were listening to TCP ports {6001, 6002, 6003, and 6004}. Moreover, random UDP background traffic continuously sent from all host in Pod 1 and Pod 2 to all hosts in pods 3 and 4, i.e., H1 to H9, H2 to H10, H3 to H11 and so forth. We used continuous UDP *iperf* requests with bandwidths ranging from 1 to 10Mbps using the default paths and values. Then a request consisting of four parallel 100Mbps TCP *iperf* of were sent from both H1 and H5 (recovery nodes) which will be handled by the servers in Pod 3 and Pod 4 respectively. Therefore, the total amount of data sent was

[4] https://github.com/Huangmachi/ECMP/blob/master/fattree4_ecmp.py
[5] https://github.com/dariobanfi/multipath-sdn-controller
[6] https://osrg.github.io/ryu/

800Mbps, and we measured the time in which all the transmissions finished.

- *Completion time*

The first metric we measured was the completion time. The results are depicted in Figure 15; as can be seen, the slowest solution was the single path approach, which completed all the data transferences in 275 seconds with a standard deviation of 16s. By contrast, the fastest one was the proposed approach in both of the cases, with and without topology augmentation, completing all the transference in a total of 30s with a standard deviation of 7.8s. In the case of the extended version of the proposal, the standard deviation was only of 4.8s which is the lowest among the benchmarked solutions. Note that ECMP completion time is also relatively low, 36.5s was the completion time of the last flow with a standard deviation of 6.4s. Finally, note that although MPSDN had a long completion time, 114s for the last flow, the variation of flow arrival was slightly higher than ours, which is due to the condition they use to ensure paths with compatible delays.

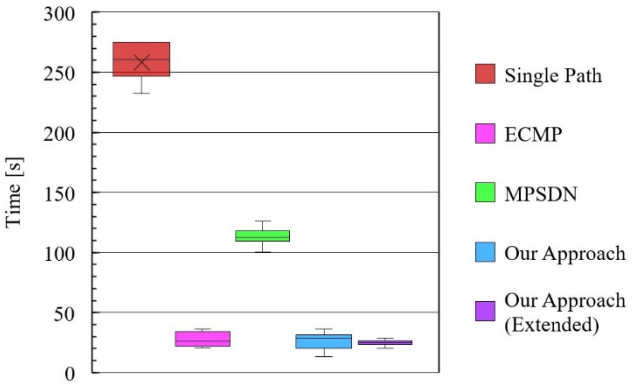


Figure 15: Completion time of 800Mbytes in a 4-ary FTN

- *Throughput*

The second metric we measured was the average throughput per flow. As shown in Figure 16, MPSDN and the simple path approach showed a steady average throughput in all the flows. However, the overall maximum performance was low, which affected the completion time. In the other cases, even though the distribution was not as regular, the overall performance was much higher. Also, it is also worth noting that since the UDP traffic was sent at random bandwidths, some of the paths were more saturated than others, which might have caused the irregular distribution. Nevertheless, as can be seen in Figure 17, in our approach, the aggregated throughput was much higher than the other approaches. In the case of the augmented version, consider that although the number of links was duplicated from the edge to the aggregation layer, the increase was not linear.

- *Network Load*

The last variable we measured was the overall network load. Although the proposed approach can handle the dynamic load balance of servers and network at the same time as shown in [25],

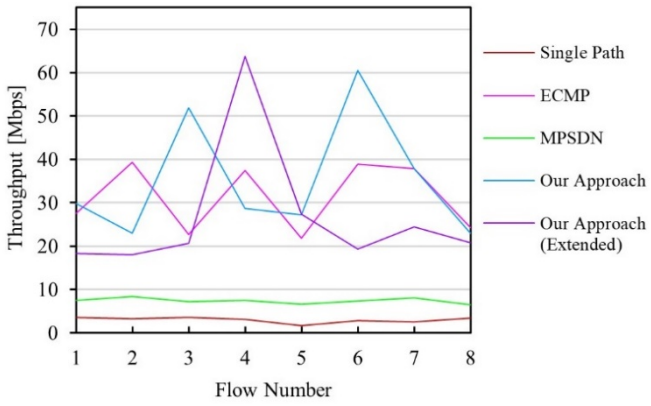


Figure 16: Average throughput per flow

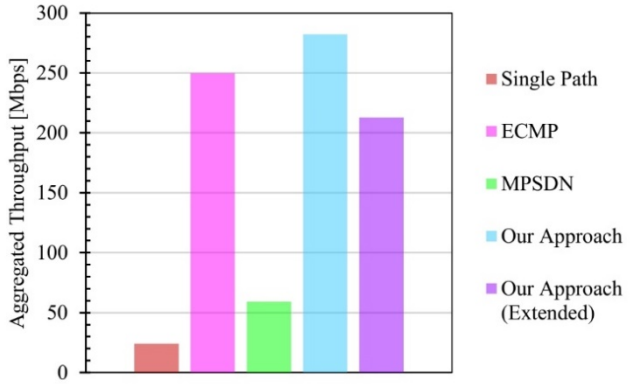


Figure 17: Aggregated throughput at eight 100Mbps flows

since the existing solutions focus mainly on the network load balancing we measured only that parameter. To perform this calculation, we used Tang's [9] formula to calculate the network bandwidth utilization ratio $\overline{\lambda(t)}$, which gives the total bandwidth utilization in a time t and is expressed as in (3)

$$\overline{\lambda(t)} = \frac{\Sigma_{1 \le i,j \le N} \lambda_{i,j}(t)}{N} \quad (3)$$

Where N is the number of links, and $\lambda_{i,j}$ is the link bandwidth utilization ratio defined as in (4), where $b_{i,j}$ is the used bandwidth and $B_{i,j}$ is the capacity of the link.

$$\lambda_{i,j}(t) = \frac{b_{i,j}}{B_{i,j}} \quad (4)$$

To make a fair comparison, we only measured the time until the time the first flow finished, which was about 30s from the beginning of the experiment. Figure 18 shows the Normalized the Cumulative Load over those 30s. As observed, ECMP was the solution that generated the most load over that time and across the whole network. However, the proposed approach used slightly more than the load of the single path solution with the added value of having a faster completion time and better overall throughput. Note that in our case, the load was not concentrated in a specific part of the network as in the other solutions, but spread among the

discovered paths. Moreover, the results for the augmented version of the topology shows better traffic distribution, which will allow more concurrent operations to be performed in less time, a condition that is desirable for DSS.

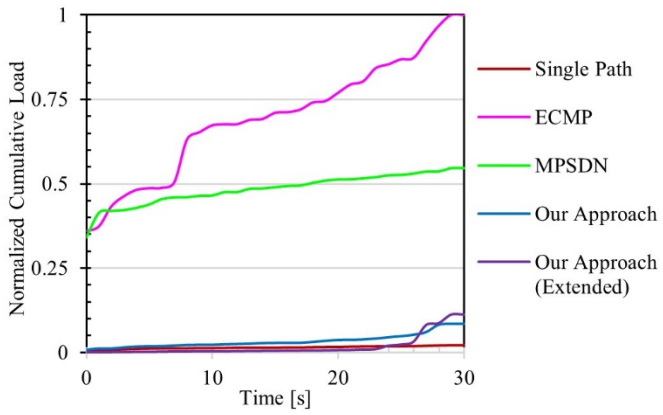


Figure 18: Normalized Cumulative Network Load in 30s

Finally, Table 4 shows a summary of the results per flow. As observed, the proposed approach outperformed the others in most of the cases. Moreover, the overall load of the entire network is compared to the one on a single path, but the traffic is more evenly distributed. Note that the network load in ECMP and MPSDN per flow is relatively high, in the case of MPSDN, it might be due to the high amount of control messages sent by the controller, but in the case of ECMP, the use of multiple paths caused several retransmissions due to packet loss.

Table 4: Average benchmark results per flow

Metric	Single Path	ECMP	MPSDN	Our App.	Our App. (extended)
Completion Time [s]	256.5	27.5	113.6	26	24.4
Throughput [Mbps]	3.02	31.24	7.41	35.28	26.61
Network Utilization [%]	0.05%	2.06%	1.45%	0.11%	0.05%

5. Conclusions

Due to the restrictiveness of using legacy network techniques for DSS, it may not be possible to cope with the ever-increasing user content-generation in the future. Therefore, in this paper, we presented a network control method to improve DSS performance by using SDN. The proposed approach used pragmatic and straightforward solutions to handle DSS's main requirements, namely aggregated bandwidth and effective use of resources. We have evaluated the proposal in various scenarios, and compared it with both network legacy techniques and existing solutions. Experimental results show that our method could achieve faster data transference and maintain balanced network and server loads. Additionally, in contrast to other proposals, our method does not require protocol or end-point modification but instead a topology enhancement based on the system demands. We showed that the overall performance of routinely DSS tasks can be dramatically increase by augmenting the DCN using parallel links, a proper path discovery, and dynamic adaption based on the network and servers load. As a future work, we still have to test the solution in large-scale and real DSS deployments. Furthermore, with the rise of new multipath transport protocols, such as QUIC [27] it might be interesting to study the impact on the way traffic is handled in non-traditional transport protocols in future networks. Likewise, we have not fully exploited the adaptive aspect of the proposal, which can be used to solve problems concerning resilient and fail-tolerant networks. Finally, it is also important to find a mechanism that achieves a tradeoff between high-performance and energy-efficiency of the whole network to cope with the proposed overprovisioned infrastructure.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] L. Guillen, S. Izumi, T. Abe, T. Suganuma, H. Muraoka, "SDN implementation of multipath discovery to improve network performance in distributed storage systems" in 13th International Conference on Network and Service Management, Japan, 2017. https://doi.org/10.23919/CNSM.2017.8256054

[2] D. Reinsel, J. Gantz, J. Rydning, "Data Age 2025: The evolution of data to life-critical - Don't focus on big data; focus on the data that's big", IDC White Paper, 1–25, 2017.

[3] A. Dimakis, P. Godfrey, Y. Wu, M. Wainwright, K. Ramchandran, "Network coding for distributed storage systems" IEEE Transactions on Information Theory, **56**(9), 4539–4551. 2010. https://doi.org/10.1109/TIT.2010.2054295

[4] C. Suh, K. Ramchandran, "Exact-Repair MDS Code Construction Using Interference Alignment" in IEEE Transactions on Information Theory, **57**(3), 1425–1442, 2011. https://doi.org/10.1109/TIT.2011.2105003

[5] N. B. Shah, K. V. Rashmi, P. V. Kumar, K. Ramchandran, "Distributed storage codes with repair-by-transfer and nonachievability of interior points on the storage-bandwidth Tradeoff" in IEEE Transactions on Information Theory, **58**(3), 1837–1852, 2012. https://doi.org/10.1109/TIT.2011.2173792

[6] Open Networking Foundation (ONF) Software-Defined Networking (SDN) Definition Available at https://www.opennetworking.org/sdn-definition/ Accessed on 2018-07-20

[7] S. Kaneko, T. Nakamura, H. Kamei, and H. Muraoka, "A Guideline for Data Placement in Heterogeneous Distributed Storage Systems" in 5th International Congress on Advanced Applied Informatics, Kumamoto Japan, 942–945, 2016. https://doi.org/10.1109/IIAI-AAI.2016.162

[8] IEEE Std. 802.3ad "IEEE Standard for Local and metropolitan area networks - Link Aggregation," available at http://www.ieee802.org/3/ad/ Accessed on 2018-06-10

[9] F. Tang, L. Yang, C. Tang, J. Li, M. Guo, "A Dynamical and Load-Balanced Flow Scheduling Approach for Big Data Centers in Clouds" IEEE Transactions on Cloud Computing, 2016. https://doi.org/10.1109/TCC.2016.2543722

[10] IETF RFC2992 "Analysis of an Equal-Cost Multi-Path Algorithm" available at https://tools.ietf.org/html/rfc2992, Accessed on 2018-01-30

[11] L. Li, Q. Xu, "Load Balancing Researches in SDN: A Survey," in 7th IEEE Int. Conference on Electronics Information and Emergency Communication, 403–408, Macau China, 2017. https://doi.org/10.1109/ICEIEC.2017.8076592

[12] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, A. Vahdat, "Hedera: Dynamic flow scheduling for data center networks," In 7th Conference on Networked Systems Design and Implementation, 19–34, San Jose CA USA, 2010. http://dl.acm.org/citation.cfm?id=1855711.1855730

[13] R. van der Pol, M. Bredel, A. Barczyk, B. Overeinder, N. van Adrichem, O. Kuipers, "Experiences with MPTCP in an Intercontinental OpenFlow Network," In 29th Trans European Research and Education Networking Conference,1–8, Amsterdam The Netherlands, 2013.

[14] IETF RFC 6182 "Architectural Guidelines for Multipath TCP Development," available at https://datatracker.ietf.org/doc/rfc6182/, Accessed on 2017-06-15

[15] J.P. Sheu, L. Liu, R.B. Jagadeesha, Y. Chang, "An Efficient Multipath Routing Algorithm for Multipath TCP in Software-Defined Networks," in European Conference Networks and Communications, 1–6, Athens Greece, 2016. https://doi.org/10.1109/EuCNC.2016.7561065

[16] D. J. Kalpana, K. Kataoka, "SFO: SubFlow Optimizer for MPTCP in SDN," in 26th Int. Telecommunication Networks and Applications Conference, Dunedin New Zealand, 2016. https://doi.org/10.1109/ATNAC.2016.7878804

[17] D. Banfi, O. Mehani, G. Jourjon, L. Schwaighofer, R. Holz, "Endpoint-transparent Multipath Transport with Software-defined Networks," in IEEE 41st Conference on Local Computer Networks, 307-315, Dubai UAE, 2016. https://doi.org/10.1109/LCN.2016.29

[18] F. Carpio, A. Engelmann, A. Jukan, "DiffFlow: Differentiating Short and Long Flows for Load Balancing in Data Center Networks," in IEEE Global Communications Conference, 1–6, Washington DC USA, 2016. https://doi.org/10.1109/GLOCOM.2016.7841733

[19] Y. Li, D. Pan, "OpenFlow based Load Balancing for Fat-Tree Networks with Multipath Support," In 12th IEEE International Conference on Communications, Budapest Hungary, 2013.

[20] S. Izumi, M. Hata , H. Takahira, M. Soylu, A. Edo,T. Abe and T. Suganuma, "A Proposal of SDN Based Disaster-Aware Smart Routing for Highly-available Information Storage Systems and Its Evaluation," International Journal of Software Science and Computational Intelligence, **9**(1), 68–82, 2017. https://doi.org/10.4018/IJSSCI.2017010105

[21] K.T. Dinh, S. Kuklinski, W. Kujawa, M. Ulaski, "MSDN-TE: Multipath Based Traffic Engineering for SDN," in 8th Asian Conference on Intelligent Information and Database Systems, 630–639, Da Nang, Vietnam 2016. https://doi.org/10.1007/978-3-662-49390-8_61

[22] P.H. Fredette, "The past, present, and future of inverse multiplexing," in IEEE Com. Magazine, **32**(4), 42–46, 1994. https://doi.org/10.1109/35.275334

[23] R. Toghraee, Learning OpenDaylight - The Art of Deploying Successful Networks, Packt Publishing Ltd., 2017.

[24] J. W. Suurballe, R. E. Tarjan "A quick method for finding shortest pairs of disjoint paths," in Networks, **14**(2), 325–336, 1984. https://doi.org/10.1002/net.3230140209

[25] L. Guillen, S. Izumi, T. Abe, T. Suganuma, H. Muraoka "SDN-based hybrid server and link load balancing in multipath distributed storage systems" in 2018 IEEE/IFIP Network Operations and Management Symposium, Taipei Taiwan, 2018. https://doi.org/10.1109/NOMS.2018.8406286

[26] M. Kuźniar, P. Perešíni, D Kostić "What You Need to Know About SDN Flow Tables," in: Mirkovic J., Liu Y. (eds) Passive and Active Measurement. PAM 2015. Lecture Notes in Computer Science, 347–359, 2015. https://doi.org/10.1007/978-3-319-15509-8_26

[27] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, et al "The QUIC Transport Protocol: Design and Internet-Scale Deployment." in Proc. of the ACM Conference of Special Interest Group on Data Communication, 183–196, New York USA, 2017. https://doi.org/10.1145/3098822.3098842

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 142-146(2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A New Study Performance Control of PMSMs: Validity Abacus Approach

Sabrine Jebri*, Khaled Nouri

Laboratoire de Systèmes Avancés (LSA), École Polytechnique de Tunisie, Université de Carthage, Tunisia

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 02 June, 2018
Online: 18 June, 2018

Keywords:
Validity Abacus Approach
NDIC
PMSM
Tracking Control
Rotor Speed
Direct Current
Frequency and Amplitude
Dependence Model

ABSTRACT

This paper is an extension of work originally presented in conference name" Study Performance of Speed Tracking Control with Frequency and Amplitude Dependence". Though many PMSMs (Permanent Magnet Synchronous Motor) control approach have already been published, little endeavors have been invested in the study of speed tacking control with frequency or amplitude dependence. The use of NDI (Nonlinear Dynamic Inversion) tracking control in versus amplitude and frequency dependence is the major focus of this paper. The proposed tracking control demonstrates that NDI control presents the ability to track the desired output with high precision. PMSM outputs are characterized at different amplitude and frequency. The speed and the direct current versus amplitude and frequency are successfully extracted. A mathematical (d-q) of the permanent magnet synchronous motor model is also presented. A comparative study between desired data and simulation results is conducted. Important to note, the proposed model and the simulation results reflect a successful agreement as far as a amplitude and frequency dependence is concerned. A new validity approach, namely "Validity Abacus" is proposed and discussed.

1. Introduction

In the electric drives systems, more than 70% of produced electrical energy is converted into mechanical energy. Therefore, a high energy efficiency of electric drives systems is demanded [1-2]. PMSMs (Permanent Magnet Synchronous Motors) gain more attention in variable speed drives due to their good efficiency, fast torque response, high power density, cooling, and lack of maintenance.

There are two types of PMSM: motors with surface mounted permanent magnets (SPMSM) and motors with internally mounted permanent magnets (IPMSM). The main difference between these two types is the existence of reluctant torque component which exists only in IPMSM due to the rotor salience [2]. The rotor of IPMSM has high mechanical strength and therefore this type of motor can operate in high speed drives [3-4].

Output tracking control has large applications in dynamic processes in industry, biology, and economics [5-6-7-8]. The most important objective of tracking control is to construct the output of the system, via a controller and track the output of a given reference model as close as possible [9].

*Corresponding Author: Sabrine Jebri, Email: jebrisabrine@yahoo.fr

In a PMSM system, there are wide quantities of the uncertainties and disturbances, unmodeled dynamics, friction force [10-11-12]. It will be very hard to limit these disturbances rapidly if adopting some classical control methods and also it is very difficult to solve the tracking control problem with these methods.

To circumvent some afore mentioned problems, an NDIC (Nonlinear Dynamic Inversion Control) was introduced, in which limit these disturbances rapidly and this approach solve the tracking control problem [1-2-13]. NDIC has been used for the study of direct current and speed tracking in PMSM, through this method, simulation results were obtained and compared. A performance study of this control technique was proposed and developed and a new validity study namely "Validity Abacus" is proposed in this work.

The present paper falls into five sections. It unfolds with a section about the PMSM control and NDIC approach. Section 2 characterize a (d-q) mathematical model of PMSMs and proposed an implementation of NDIC approach of the direct current and the speed tracking control of PMSM. Section 3 introduces the frequency and amplitude dependent model. Simulation results are discussed and conclusions comments are given respectively in sections 4 and 5.

2. Preliminaries

In this section, some preliminaries are introduced, including PMSM (d-q) model and NDIC method, to propose a new tracking control approach.

2.1. Proposed Model

The important things are the design of tracking control systems to realize the variation speed and high precision motion. The PMSM electrical equations can be written in the (d-q) rotor flux reference [14-15] as following:

$$\dot{I}_d = -\frac{R_s}{L_d} I_d + \frac{pL_q}{L_d} I_q \Omega + \frac{1}{L_d} u_d \quad (1)$$

$$\dot{I}_q = -\frac{R_s}{L_q} I_q - \frac{pL_q}{L_d} I_d \Omega - \frac{p\,\phi}{L_q} \Omega + \frac{1}{L_q} u_q \quad (2)$$

$$\dot{\Omega} = \frac{p\,\phi}{J} I_q - \frac{\mathrm{B}}{J} \Omega + \frac{1}{J} C_T \quad (3)$$

I_d and I_q: d and q axis stator currents.

Ω: rotor speed.

U_d and U_q: d and q stator voltages

2.2. NDIC Proposed Method

In this paper, the NDIC proposed approach will be used as a tracking controller for the direct current and the rotor speed of PMSMs to investigate its utility [1-16-17-18]. Figure 1 present the structure of the inverting law of the nonlinear dynamic inversion control approach.

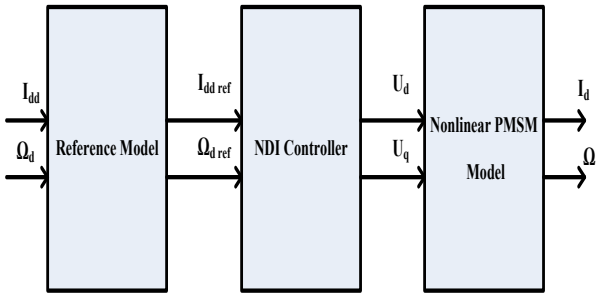


Figure.1. Nonlinear Dynamic Inversion Control approach [1]

The control objective is to track a desired direct current (I_{dd}) and speed (Ω_d) with respectively the actual direct current (Id) and the rotor speed (Ω). The dynamic nonlinear model of the PMSM expressed in equations (1, 2 and 3) is taken in the following form:

$$\begin{cases} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{bmatrix} = \begin{bmatrix} \dot{I}_d \\ \dot{I}_q \\ \dot{\Omega} \end{bmatrix} = \begin{bmatrix} \dot{I}_d = -\frac{R_s}{L_d} I_d + \frac{pL_q}{L_d} I_q \Omega \\ \dot{I}_q = \frac{R_s}{L_q} I_q - \frac{pL_q}{L_d} I_d \Omega - \frac{p\,\phi}{L_q} \Omega \\ \dot{\Omega} = \frac{p\,\phi}{J} I_q - \frac{\mathrm{B}}{J} \Omega + \frac{1}{J} C_T \end{bmatrix} + \begin{bmatrix} \frac{1}{L_d} & 0 \\ 0 & \frac{1}{L_q} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_d(t) \\ U_q(t) \end{bmatrix} \\ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} I_d \\ \Omega \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} I_d \\ I_q \\ \Omega \end{bmatrix} \end{cases} \quad (4)$$

With:

$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} I_d \\ I_q \\ \Omega \end{bmatrix}$: is the state vector system

$\begin{bmatrix} U_d(t) \\ U_q(t) \end{bmatrix}$: is the input vector

$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} I_d \\ \Omega \end{bmatrix}$: is the output vector

$$I_{dd}(t) = \mathrm{I}_{max} \sin(t^2) \quad t \in [0, 2\pi] \quad (5)$$

$$\Omega_d(t) = \Omega_{\max}(1 - \exp(2t)) \quad t \in R \quad (6)$$

Using the nonlinear dynamic inversion proposed approach; the input can be obtained using the successive derived from the output. Then $U_d(t)$ is expressed by equation 7.

$$u_d(t) = \frac{1}{\beta_1(t)} \left[\dot{I}_{dd}(t) - \alpha_1(t) \right] \quad (7)$$

With:

$$\alpha_1(t) = -\frac{R_s}{L_d} I_d + \frac{pL_q}{L_d} I_q \Omega_d \quad (8)$$

and

$$\beta_1(t) = \frac{1}{L_d} \quad (9)$$

The input $U_q(t)$ is expressed by equation (10):

$$u_q(t) = \frac{1}{\beta_2(t)} \left[\ddot{\Omega}_2(t) - \alpha_2(t) - \frac{\mathrm{B}}{J^2} C_T \right] \quad (10)$$

With:

$$\alpha_2(t) = \frac{p\,\phi}{J} \left(\frac{\mathrm{B}}{J} - \frac{R_s}{L_q} \right) I_q - \left(\left(\frac{\mathrm{B}}{J} \right)^2 - \frac{(p\,\phi)^2}{JL_q} \right) \Omega - \left(\frac{p^2\,\phi L_q}{JL_d} \right) I_d \Omega \quad (11)$$

and

$$\beta_2(t) = \frac{p\,\mathrm{f}}{JL_q} \quad (12)$$

3. Validity Abacus Approach

In order to validate the proposed nonlinear dynamic inversion approach, the following methodology is used in figure 2.

#first step: the mathematical nonlinear (d-q) model of PMSM is developed.

#Second step: this step consists in describing the frequency dependence of the speed tracking control model. In the third step, simulation results are compared to the desired model for both the direct current I_{dd} and the speed Ω_d.

 # Third step: the errors between the desired data and the simulation results are computed. Based on these errors, discrepancies between them are depicted in order to finally judge on the correction of the model parameters. Figure 2 summarizes the previously presented steps respectively.

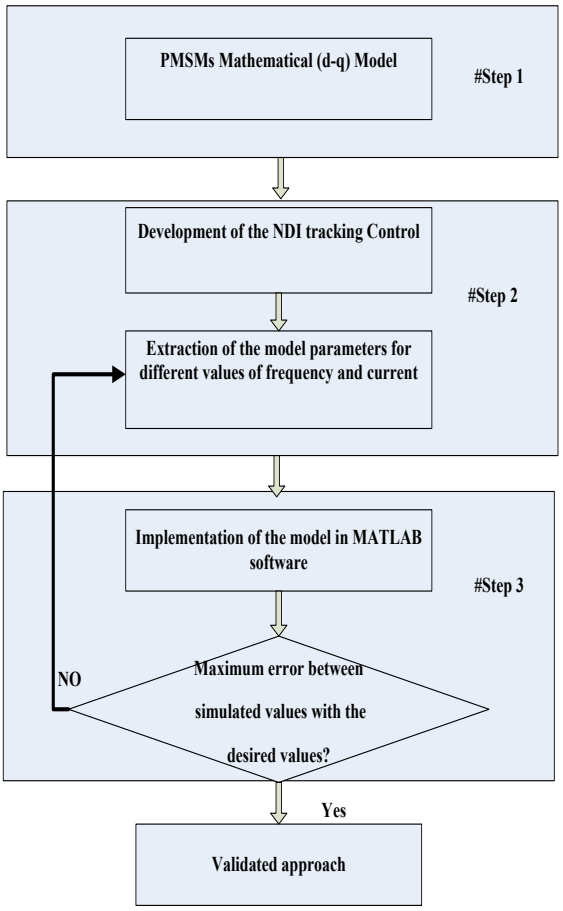


Fig.ure 2.Validity Abacus Methodology

Important to note, that the estimated errors between the desired outputs and the obtained ones, are in good agreements. This paper presents.

A novel to easy methodology to easy study performance of control methods by only optimization error between desired and simulated results for rotor speed and direct current at different frequency and amplitude levels. This methodology may be applied for others control approach.

Table 1: PMSM parameters and values [1]

Parameter	Denotation	Value
P	Pole pair number	2
R_s	Stator resistance	1.9 Ω
L_d	d-axis inductance	42.44 mH
L_q	q-axis inductance,	79.57 mH
f	Frequency	50 Hz
φ_f	Magnetic flux constant,	0.311 Wb
J	Inertia of the motor	0.003 kg.m2
C_T	Motor torque	3.954 Nm
B	Friction coefficient	0.001 Nm/rad/s

4. Simulation Results

The rotor speed and direct current tracking control using NDIC approach and the frequency and amplitude dependent model of PMSM are emulated in MATLAB. Table I present the numerical parameters of the studied model.

Figure 3 and figure 4 shows the speed characteristics at different frequency levels respectively in low and high frequency. The different amplitude of the desired direct current (see figure 5) the change in the saturation speed.

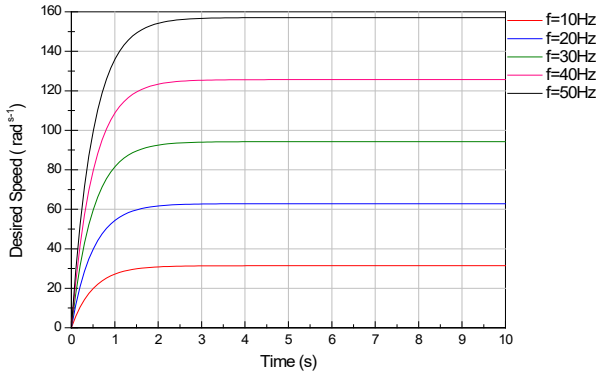


Figure.3. Desired speed characteristics at different frequency (low- frequency f= [10 ➔50Hz]).

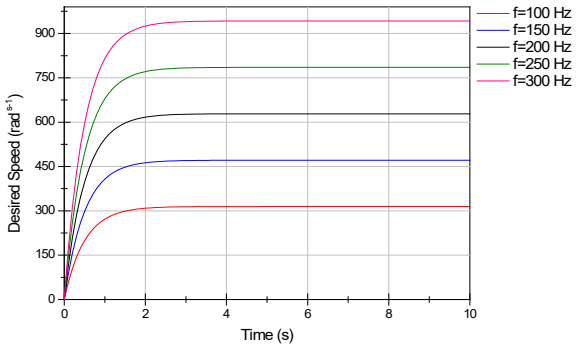
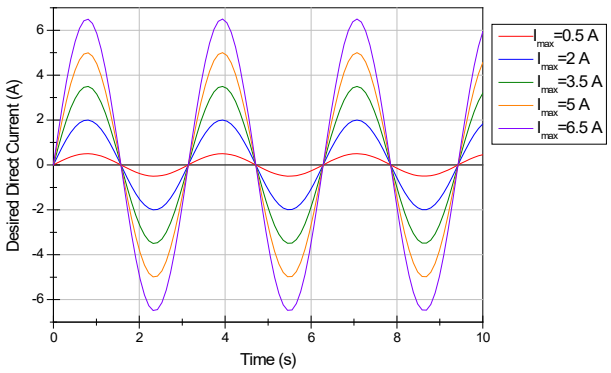
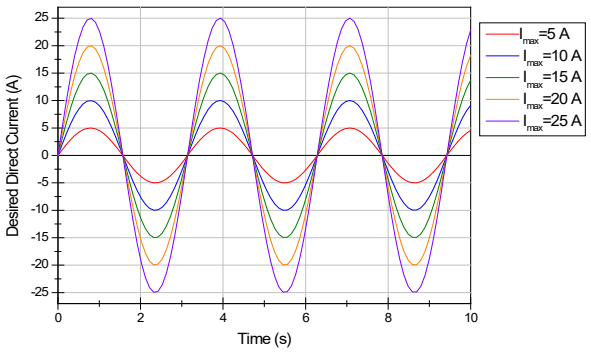


Figure.4. Desired speed characteristics at different frequency (high- frequency f= [100 ➔300Hz]).

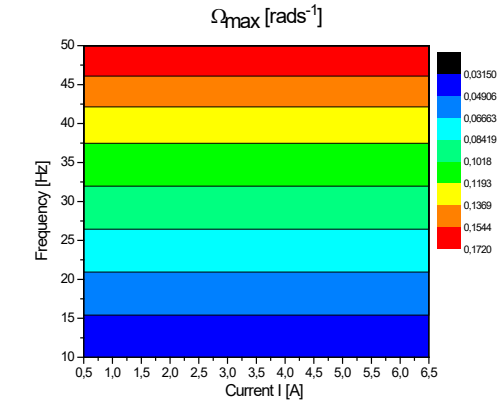


(a)

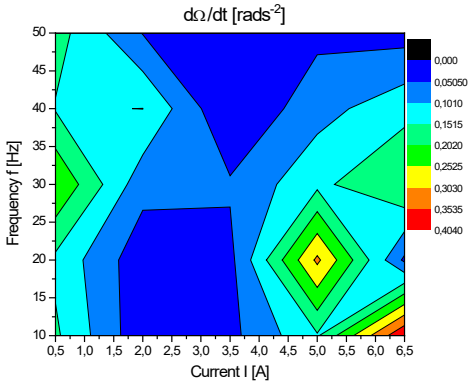


(b)

Figure 5. Desired current characteristics at different amplitude
(a). I *max*= [0.5➔6.5A] for the low-frequency study and
(b).I*max*= [5➔25A] for the high-frequency study



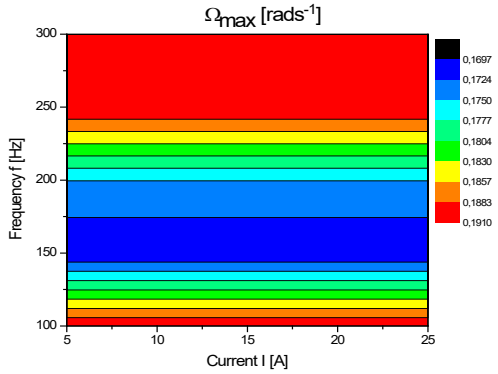
(a)



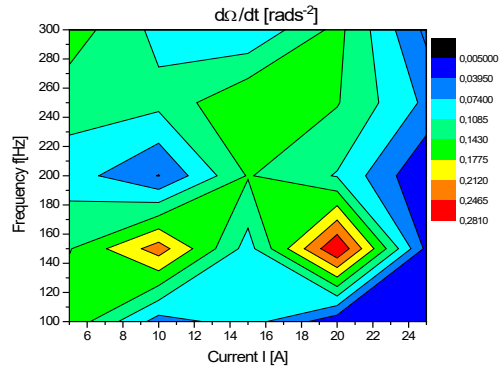
(b)

Figure.6. Validity abacus of rotor speed tracking control in low- frequency: (a) Ω, (b) dΩ/dt

Figures 6 and 7 are an illustration of validity abacus of tracking parameter error (Ω, dΩ/dt). The rotor speed tracking of permanent magnet synchronous motor has been performed for different data both in low and high frequency. These latter, may be effective in the optimization of rotor speed tracking using a nonlinear dynamic inversion control approach. Then, these abacuses evaluate the accuracy of simulation results.



(a)



(b)

Figure.7. Validity abacus of speed tracking control in high-frequency
(a) Ω, (b) dΩ/dt

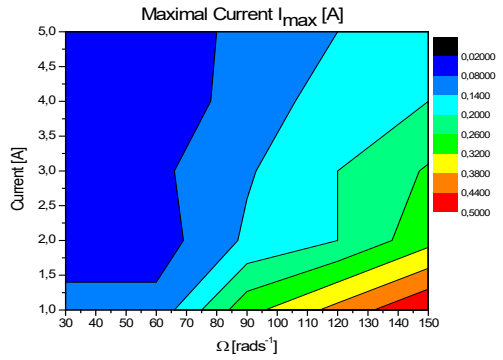


Fig.ure 8. Validity abacus of direct current for rotor speed [30 rad/s ➔150rad/s]

The Ω abacus (Figure 6.a and 7.a) illustrates that the nonlinear dynamic inversion control approach of the rotor speed tracking PMSM parameters offers a wide range of validity. The error between the estimated maximal speed rotor Ω_d and the obtained simulation is less than 0.2% for different rotor speed level in low and high frequency. The dΩ/dt validity abacus shows an error less than 0.5% at low and high frequency (Figure 6.b and 7.b). The I_d abacus (Figure 8) illustrate that NDIC of the direct current tracking offer a wide range of validity. The error between the estimated maximal direct current I_{dd} and the obtained simulation get by this approach is less than 1% for different current level for both the proposed PMSM.

The error is mainly due to the nonlinearity of PMSM model equations. Parameters error is less than 0.5% for different conditions. The obtained result proves the effectiveness of the control to ensure a rotor speed and direct current tracking. The abacus also helps the engineer to evaluate the simulation results accuracy of given devices. Various upshots and diverse validity abacus are satisfying so far. For a low and high frequency conditions, the proposed methodology is then adequate. Moreover, this model offers an efficient trade-off between ease and accuracy of speed tracking control study.

However, simulation errors cannot be avoided. They are calculated according to the rotor speed and the direct current at different frequency and amplitude. An error less than 0.5% is found for the permanent magnet synchronous motor model proposed. In addition to its precision, the suggested model has got a good convergence. Furthermore, the ease of use is one more advantage. It consists in both the facility of implementation using MATLAB and the simplicity in extracting PMSM model parameters. Moreover, the proposed algorithm, integrating the developed model, allows a better detection and adjustment of the outputs tracking control (using NDIC approach). In fact, in the case of poor design, the model replies routinely by detecting and tuning it. Nonlinear dynamic problems are, therefore, methodically avoided.

5. Conclusion

In this paper we have demonstrated that the PMSMs have the direct current and the rotor speed proportional to the frequency and amplitude. A mathematical nonlinear model of this motor is also proposed. The novel of this model is that it takes into account the design structure as well as frequency and amplitude parameters. The model validation is performed by comparing desired results with simulation. This model is run for a large range frequency and amplitude, and the simulation investigations show a good agreement with respect to the desired ones. These results are validated and. Future works concern the validation of the NDIC proposed approach associate the “Abacus Validity” approach others nonlinear systems.

References

[1] S. Jebri , K. Nouri and L. Bouslimi,” Study Performance of Speed Tracking Control with Frequency an Amplitude Dependence”IC_ASET, IEEE conferences Tunisia, Janvier 2017.

[2] S. Jebri ; H. Loussifi ; K. Nouri,” Two Mass Modeling and Nonlinear Dynamic Inversion of Motor Drive Systems”. Sciences and Techniques of Automatic Control and Computer Engineering, IEEE conf . July 2016.

[3] J. Hu, D. Dawson, and Y. Qian, “Position tracking control of an inductionmotor via partial state-feedback,” Automatica, vol. 31, pp. 989–1000, 1995.

[4] D. M. Dawson, Z. Qu, and J. J. Carroll, “Tracking control of rigid-link electrically-driven robot manipulators,” Int. J. Control, vol. 56, no. 5, pp. 991–1006, 1992.

[5] Z. Qu and J. Dorsey, “Robust tracking control of robots by a linear feedback law,” IEEE Trans. Autom. Control, vol. 36, no. 9, pp. 1081–1084, Sep. 1991.

[6] L. Benvenuti, M. D. Di Benedetto, and J. W. Grizzle, “Approximate output tracking for nonlinear nonminimum phase systems with an application to flight control,” Int. J. Robust Nonlinear Control, vol. 4, no. 3, pp. 397–414, 1994.

[7] F. Liao, J. L. Wang, and G. H. Yang, “Reliable robust flight tracking control: An LMI approach,” IEEE Trans. Control Syst. Technol., vol. 10, no. 1, pp. 76–89, Jan. 2002.

[8] N. C. Shieh, P. C. Tung, and C. L. Lin, “Robust output tracking control of a linear brushless DC motor with time-varying disturbances,” Inst. Elect. Eng. Proc. Elect. Power Appl., vol. 149, no. 1, pp. 39–45, 2002.

[9] G. Bartolini, A. Ferrara, and E. Usai, “Output tracking control of uncertain nonlinear second-order systems,” Automatica, vol. 33, no. 12, pp. 2203– 2212, 1997.

[10] X. Chen and C.-Y. Su, “Robust output tracking control for the systems with uncertainties,” Int. J. Syst. Sci., vol. 33, no. 4, pp. 247–257, 2002.

[11] P. Seiler and R. Sengupta, “An H∞ approach to networked control,” IEEE Trans. Autom. Control, vol. 50, no. 3, pp. 356–364, Mar. 2005.

[12] N. C. Shieh, K. Liang, and C. Mao, “Robust output tracking control of an uncertain linear system via a modified optimal linear-quadratic method,” J. Optim. Theory Appl., vol. 117, no. 3, pp. 649–659, 2003.

[13] D.Yang and J. Zhao,” H1 output tracking control for a class of switched LPV systems and its application to an aero-engine model” .International Journal of Robust and Nonlinear Control, September 2016.

[14] Y.Xin Li and G.Hong Yang,” Fuzzy Adaptive Output Feedback Fault-Tolerant Tracking Control of a Class of Uncertain Nonlinear Systems With Nonaffine Nonlinear Faults”IEEE Transactions on Fuzzy Systems .Vol.24, no: 1, Feb. 2016.

[15] B.A. Francis and W.M. “Wonham. The internal model principle of control theory”. Automatica, 12(5–E):457–465, 1976.

[16] M. Koç, J. Wang, T. Sun,” An Inverter Nonlinearity Independent Flux Observer for Direct Torque Controlled High Performance Interior Permanent Magnet Brushless AC Drives” IEEE Trans. Power Electron, Volume:PP , Issue: 99, February 2016.

[17] T. D. Do, S. Kwak, H. H. Choi, and J. W. Jung, “Suboptimal control scheme design for interior permanent magnet synchronous motors: An SDRE-based approach,” IEEE Trans. Power Electron., vol. 29, no. 6, pp. 3020–3031, Jun. 2014.

[18] K. Nouri, R. Dhaouadi, and N. B. Braiek, “A dynamic neural network based decoupled extended Kalman filter control applying to motor drive system,” 17ème Congrés Mondial IMACS, Calcul Scientifique, Mathématique Appliquées et Simulation, 2005

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 147-151 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Automating Hostel Telephone Systems

Rohan Prabhu Murje, Bhaskar Rishab, Krishna Gopalrao Jorapur, MuccatiraThimmaiah Karumbaiah, Muddenahalli Nagendrappa Thippeswamy*

Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Yelahanka, Bengaluru, India-560064

ARTICLE INFO

Article history:
Received: 22 April, 2018
Accepted: 01 June, 2018
Online: 18 June, 2018

Keywords:
Communication
Digital lines
EPABX systems
PRI cards

ABSTRACT

Residential schools can be a great place for students to learn and develop a different set of skills if the right environment is given to the child. Constant communication with parents can help the child to stay motivated and feel comfortable away from home, hence schools tend to provide various solutions to allow the student to communicate with their parents. The most common ways are using a coin booth or a common telephone. This process is very hectic as hundreds of student's use this facility and these hundreds of calls have to be monitored individually so that no student would be getting into bad influence. In order to counter this problem, the entire process is automated in this paper. The proposed work also countered every small problem, such as unauthorized access, making calls to unauthorized numbers or exceeding the duration of call time allotted. In this paper, the digital lines are used to allow multiple calls and control these with the help of a Primary Rate Interface (PRI) card and an Electronic Private Automatic Branch Exchange (EPABX) system.

1. Introduction

The business phone systems are more affordable due to the rise of cloud technology. This has transferred the telephone industry standards [1]. Automated phone systems can manage a high volume of callers and provide a method for filtering calls in a timely manner for small business [2]. The automation can save a substantial amount of time and money [2]. The number of parents opting for a residential school for their kids is on the rise, parents expect the right atmosphere for the students to develop. The schools do their best to provide the best facilities but often fail to monitor and control the usage of these facilities. One of the most common problems faced by the institutions is manually monitoring telephone facilities provided to the students. Parents often complain about students not contacting them on a regular basis. There are many other problems, for example, not everyone gets a chance to call; unlimited talk time calls are made to unauthorized people, and unauthorized access to the facility [3]. In order to counter all these problems, an automation solution which revolves around the automation of controlling and monitoring the calls that are made is proposed. It has pre-set numbers, pre-set call time, automated duration control and many other features which help in reducing the amount of human intervention required. This is done by setting up a telephony network comprising of PRI lines, an interface card, an EPABX system and analogue phones. The back-end software controls the communication flow allowing us to manage and monitor every aspect, with minimal human intervention.

2. Related Work

The automatic voice responding system that uses the computer stored data is proposed in [4]. When a parent dialed the hostel phone number, they will get the answer in automatic stored voice form in this work. In [5], when a parent dialed the specified mobile number, they will get the student percentage attendance in voice form. Thus, parents can get their ward performance report from anywhere at any time.

As the rate at which residential institutions are growing, it is found that there is a greater need to automate the telephone calling system. It is necessary for a student to be in constant touch with his/her parents to remain motivated. The ability to quickly and efficiently manage hundreds of simultaneous calls is necessary in order to effectively manage the process [6].

*Thippeswamy M.N., Department of Computer Science and Engineering, NitteMeenakshi Institute of Technology, Bengaluru, India-560064, 00919686329815, Email: thippeswamy.mn@nmit.ac.in

In this paper, different methods are explored that have been used over the years to provide this facility. In order to facilitate the calling process, one approach is to use coin phones. Here, each student can call their respective parents by inserting coins into the phone and making calls. This isn't feasible as there can be no control and monitoring of the system as the number of students increase. The other method that has been implemented is to allow only incoming calls using an analogue line connected to an analogue phone, the incoming call can be received by the student and the facility can be used.

The provision to block the call during the closed hours is provided, avoiding unauthorized access to the facilities. Though this system can address a lot of problems it cannot address some major issues like monitoring incoming calls, controlling the access and since analogue lines are used, only one call can be made/received at a time; hence, giving rise to a new problem of investment on analogue lines and cabling. Similar telephony systems are used for intra-communications within organizations, they use similar methods to allow intercom.

An EPABX system [7] with advance feature allows staff members to monitor the telephone usage and manage calls over a large organization. In this paper, analogue line is used instead of PRI lines [8]. Such systems are also used in a Call center where an Omni channel solution is required.

Compared to the existing system of telephone management, aproposed system uses PRI lines and a PRI interface card which is connected to the system, where the server controls the needs of the institution whereas generally there is no involvement of a local server providing this functionality. The systems are only used for intercom and other such facilities hence functions like pre-assigning call timings are not present.

3. Novel Contribution

This solution allows monitoring and control of a student's telephone activities in the following ways:

i. Pre-assigned call timings, the lines only open during this time, hence avoiding unauthorized access.
ii. Pre-assigned Phone numbers, Speed dial is set to these telephone numbers, hence avoiding students calling other unknown numbers.
iii. Pre-assigned authorisation code, each student has an authorization code, which activates their account.
iv. Pre-assigned call duration, flexible call duration, admin can set call duration.
v. Automatic report generation, reports are sent every day to the administrator's mail ID.
vi. Reports based on status of the call, reports based on call status, i.e. if a call is connected or dropped, can be generated.
vii. History of call log, complete call history of the student can be checked with few clicks.
viii. Flexible call charges, the management has the privileges to charge for the facility provided.

4. System Design

The system consists of different telephony hardware, this is put-together with backend software coded in asterisks and Hypertext Pre-processor (PHP). Also the system consists of database tables created in My Structured Query Language (MySQL). The main components of the system are PRI lines, PRI card, EPABX system and a local server. The telephone systems architecture consists of various hardware devices.

This software is built on controlling the PRI lines which can allow up to 30 simultaneous calls; this digital line is connected to the local server which consists of a PRI card. This card allows the PRI lines to be connected to the server (refer to Figure 7). In this paper, for the outgoing lines, an EPABX is used, which is a branch exchange device that helps to channel the outgoing calls. PRI line is also known as a T1 line, it is a multi-channel line. The feature of a call transferring and forwarding is another area enabling mobility of the users. Auto conferencing and automatic redialling of numbers found engaged on the first trial are some of the other advancements in the features of the EPBAX. An EPABX also checks the traffic and is also called as a switching system.

4.1. A brief functionality of each component:

- ***PRI Lines:*** PRI lines or primary interface lines are digital lines, which are a telecommunication standard that enables traditional phone lines to carry voice, data and video traffic, among others. A PRI line physically has only one line terminating on the interface but can allow us to send or receive 30 calls simultaneously. A PRI circuit consists of two pairs of copper lines terminating on a modem from a service provider premises to the customer premises. It uses multiplexing/de-multiplexing techniques to carry more than one channel in a single circuit. There are two common types of PRI lines used namely E1 and T1. The E1 type consists of 30 channels used mostly in India and Europe. Type T1 consists of 23/24 channels. T1 type is mostly used in the US. In this paper, the PRI E1 type is used.
- ***PRI interface cards:*** A PRI Card is used to connect PRI lines to IP PBX/ IP Telephony Server so that the entire Analog phones (extensions) can make outgoing calls or receive incoming calls using it. This is inserted in a local server in the Peripheral Component Interconnect (PCI) 3.3V/ PCI 5V/ PCI Express slots. These cards can have one slot, two slots or four slots.
- ***EPABX System:***An EPABX system which is a branch exchange device that helps channeling the outgoing calls. It also controls the traffic and is also called a switching system.
- ***Analog Phones:***The analogue phones allow speed dialing, hence it can serve the purpose of not allowing dialing of numbers and pre-assigning telephone numbers.
- ***Local Server:***The local server consists of the software which controls communication. This software has the following modules.

- ***Student details:***This module contains all the data of the students, it contains a database which stores the name, id, auth code and three pre-assigned telephone numbers.
- ***Verification module:*** This module contains the logic which verifies the authentication of the student. The number entered on the analog phone is matched with the authentication code present in the database; the student details present under the entered auth code is loaded onto the speed dial of the telephone.
- ***Balance account:*** Here all the students account balance is checked to see if an outgoing call can be made, if not then a voice over message is played alerting the user to recharge their account. After the call terminates the latest balance is updated to the database.
- ***Timings module:*** This module contains the assigned calling time. When the phone is accessed the timing, module is accessed and the time allotted is matched with the current time, if the there is a match between current time and opening time the user is allowed to use the facility else a voice over message alerts the user about the line being closed.

- **System module:**The system architecture consists of the above mentioned hardware. First, the PRI line is terminated at a Local Area Network (LAN) switch and a LAN cable is drawn to the local server which has a PRI card. In the local servers various modules are present. From the local server a branch exchange is used and then physical wiring is done to each room of the hostel which has a telephone.

5. Implementation and Results

The solution has been implemented using PHP as a front-end for the user interface and asterisks coding is used as a back-end to control the communication process. The implementation is performed by setting up telephony hardware as follows: the PRI line is terminated at the PRI interface card, which is present in our local server. In the local server the software to control and monitor the system is locally hosted. There are LAN connections between the local server and the EPABX system, from the EPABX system there are physical wiring to the hostel room where the phones are present.

This section shows the results that have been obtained from the implementation of this solution. The Figure (s) 1 – 6 are the results obtained for the proposed system.

The User Interface (UI) of the webpage is shown in Figure 1, which prompts the admin to login to the system. Upon successful login the admin can access different features present like the student details page depicted in Figure 2, here an authorization code is given to each student and their outgoing telephone numbers are assigned.

The next module is the timings module which is shown in Figure 3, here the admin can allot timings for outgoing and incoming calls for everyday of the week.

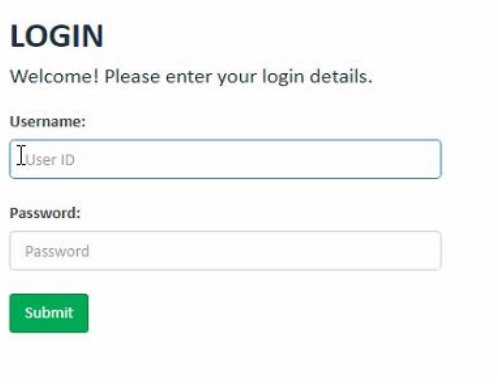


Figure 1: Shows the login page.

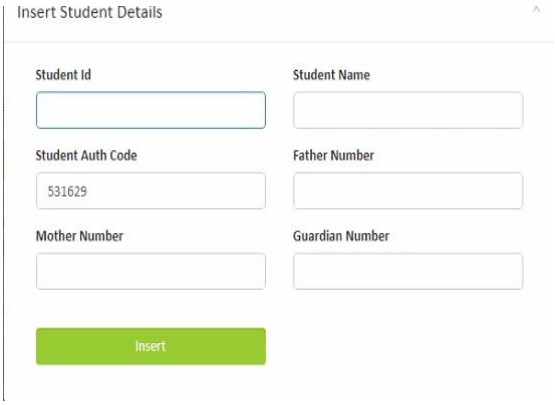


Figure 2: Represents the screen for entering student details.

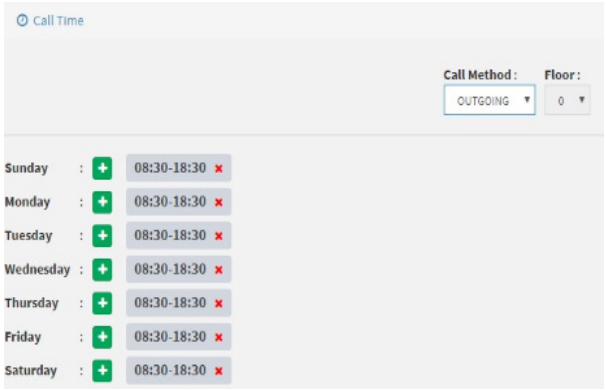


Figure 3: Shows the call timing module.

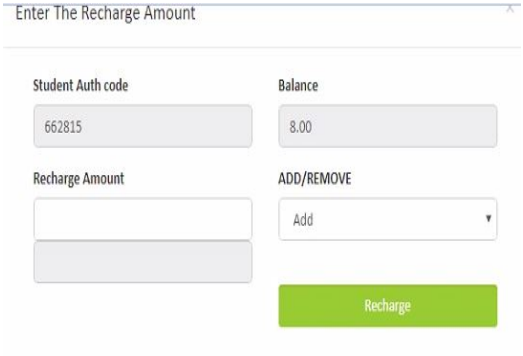
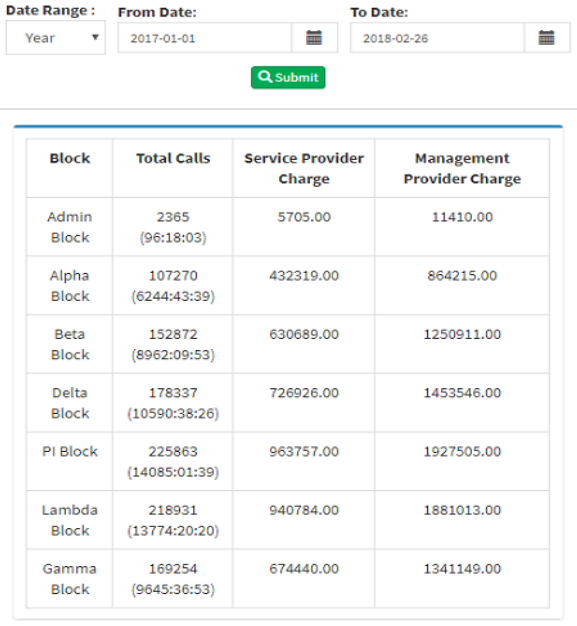


Figure 4: Shows the recharge module of the software

To make an outgoing call the admin has to recharge the students account, this is depicted in Figure 4. Once an outgoing call is made and is disconnected the database is updated and reports are generated as seen in Figure 5 (a) and (b). These bills are generated based on the flexible call charges module, which allows the management to charge over the existing price, this module is depicted in Figure 6.



Block	Total Calls	Service Provider Charge	Management Provider Charge
Admin Block	2365 (96:18:03)	5705.00	11410.00
Alpha Block	107270 (6244:43:39)	432319.00	864215.00
Beta Block	152872 (8962:09:53)	630689.00	1250911.00
Delta Block	178337 (10590:38:26)	726926.00	1453546.00
PI Block	225863 (14085:01:39)	963757.00	1927505.00
Lambda Block	218931 (13774:20:20)	940784.00	1881013.00
Gamma Block	169254 (9645:36:53)	674440.00	1341149.00

(a)

Total Calls - (1054892)	Total Duration - (63398:48:53)
Service Provider Charges	4374620
Our Charges	8729749
Our Profits	4355129

(b)

Figure 5: Shows the billing reports over a period of time.

6. Conclusion

This paper explains the working of a telephone system, which can help an institution tremendously by automating the control and monitoring of the process. This reduces the human effort that is required to monitor each and every call made by hundreds of students on a daily basis. This solution saves a lot of time and the effort required to monitor each and every call of hundreds of students. Reports are generated every day based on the status of the call, last call charge, total profits giving the management a bird's eye view of the process. Adding to the ease of control and monitoring it also gives the institution flexibility to charge per unit call for example, if the call provider charges Indian Rupees (INR) 1.0 per minute, the institution has options to charge INR 1.5 per minute, or INR 2.0 per minute or any amount they would want to. This solution would be beneficial for institutions with greater than hundred students who have taken residential facility from the school.

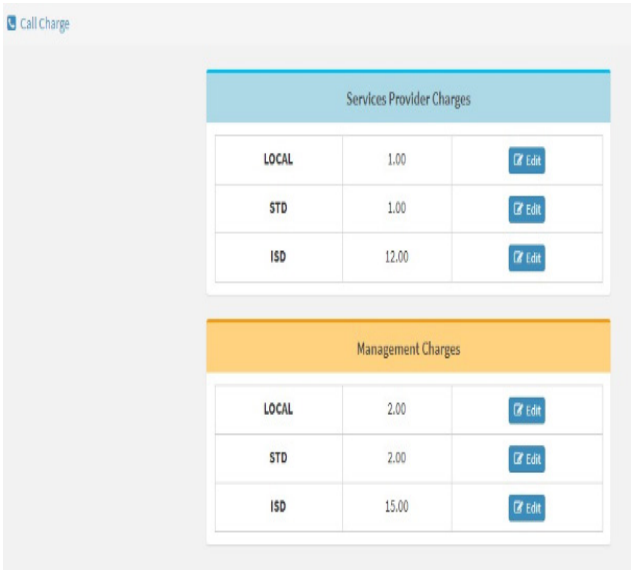


Figure 6: Shows the call charges module.

Conflict of Interest

The authors declare no conflict of interest.

References

[1]. BT Prepare to Switch to an All IP Network by 2025, [online] Available: http://www.sa1solutions.com/blog-bt-moving-ip-network.php.

[2]. A. Smith, Automated Phone Systems: Identifying the Problem with yours and how to turn the ship around, [Online] Available: https://www.theivrvoice.com/automated-phone-systems/

[3]. D. Kirk, How much is intelligent automation saving you, [online] Available:https://www.forbes.com.

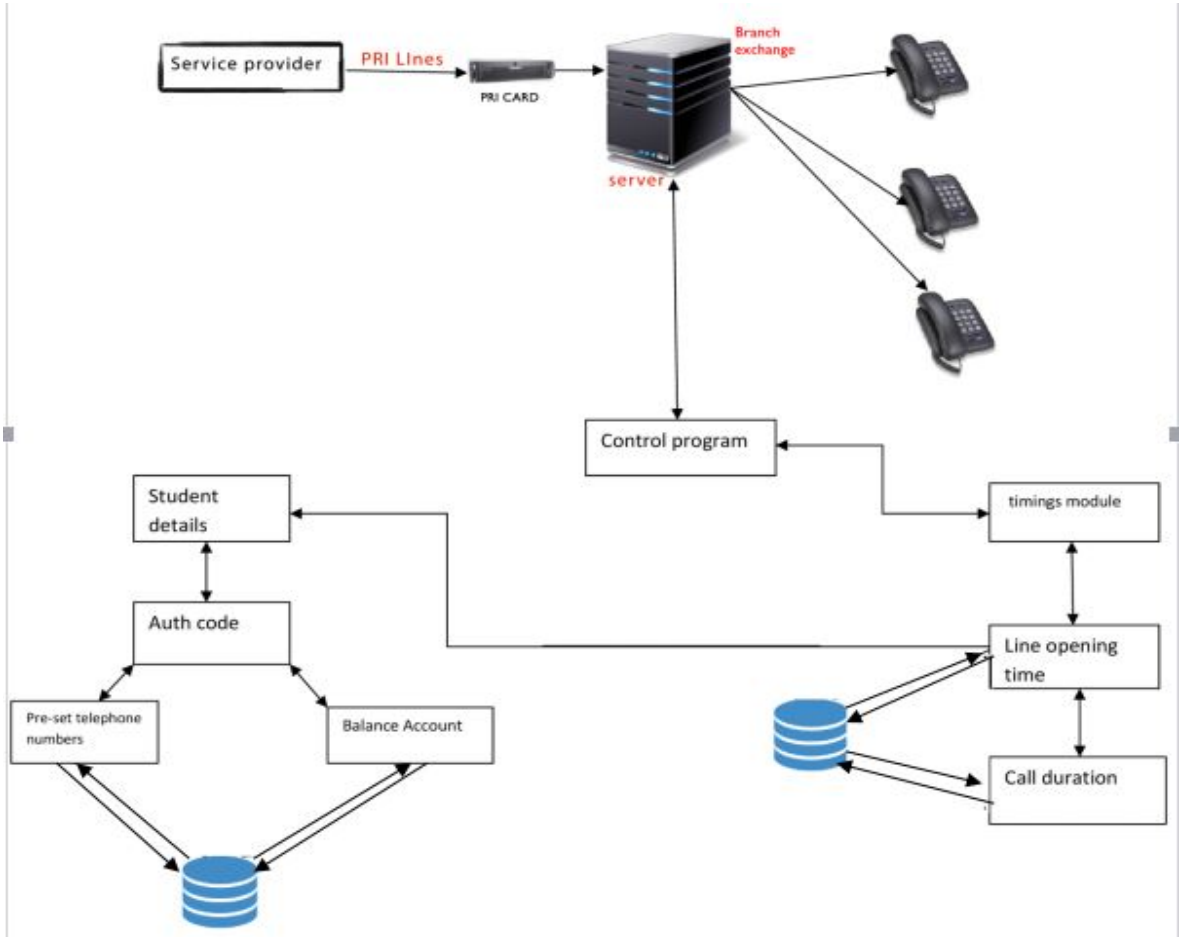
[4]. I.J.Vinila, "Voice Response System for Parents of Hostel Students", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 5, May 2016.DOI: 10.15680/IJIRCCE.2016. 0405022

[5]. J. Patil and J. Barole, "Automatic Voice Responding System for Parents of Students", International Journal of Innovative Research in Science, Engineering and Technology,Vol. 4, Issue 6, June 2015.

[6]. Hostel Telephone System, [online] Available:http://mipsoft.com/telephone.html

[7]. U. Nadeem. "The Functions of a PBX", [Online] Available: https://www.lifewire.com/functions-of-a-pbx-3426325

[8]. K. Rajesh, Telecom Trunk Lines: Difference Between Analog Line PRI/E1/T1 Digital-Line-GSM-Gateway, [online] Available: http://www.excitingip.com/2111/telecom-trunk-lines-difference-between-analog-line-prie1t1-digital-line-gsm-gateway/


Service provider
PRI LInes
PRI CARD
Branch exchange
server
Control program
Student details
Auth code
Pre-set telephone numbers
Balance Account
timings module
Line opening time
Call duration

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 152-160 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Evolving AL-FEC Application on 5G NGMN-Edge Computing Systems

Christos Bouras[*1,2], Nikolaos Kanakis[2]

[1]*Computer Technology Institute and Press "Diophantus", Patras, Greece*

[2]*Computer Engineering and Informatics Department, University of Patras, Greece*

ARTICLEINFO

Article history:
Received: 22 July, 2018
Accepted: 16 September, 2018
Online: 29 September, 2018

Keywords:
Forward error correction
Next generation mobile network
Online algorithms
Mobile edge computing

ABSTRACT

Fifth generation of mobile networks (5G) comes to cope with the context of the new mobile telecommunications era. Edge computing is a new collaborative technology under standardization utilizing end-user devices or near-end-user edge devices to operate processing, communication, or control operations. The achievement of efficient error control is a very critical aspect in the successful development of 5G Next Generation Mobile Networks (NGMN). Especially utilizing Forward Error Correction (FEC) codes on the application level is an efficient approach on improving error control on NGMN since it is adopted in several mobile multicast standards. FEC is a feedback free error control technique where redundant data are introduced with the source data to enable the recipient recovering from packet losses with an interesting approach on applying AL-FEC error protection introduces deterministic or randomized online algorithms. Based on this we present a novel online scheme on applying AL-FEC application on the context of the AL-FEC policy online problem. In this work we present an online algorithm based on feedback received from mobile end-users which adapts the introduced protection based on this. Another aspect of this work is the utilization of AL-FEC protection on the edge level based on RaptorQ FEC codes.

1 Introduction

The design of an algorithm capable to adapt the transmission redundancy that will be introduced by AL-FEC protection can be addressed with online algorithms. Online algorithms [1] are utilized on problems where the algorithm does not have knowledge of the input in advance. Hence, online algorithms should be able to make a decision without a priori knowledge of the input since it is not available at present. On online algorithms design it is assumed that algorithm's input arrives in steps and the online algorithm has to make a decision upon the arrival of each new input. Moreover, in problems where deterministic algoithms are not suitable or cannot even apply the simplest and several times the most efficient approach is utilizing randomized online algorithms [2]. Regarding the performance of online algorithms it is analyzed utilizing competitive analysis where the output produced by the evaluated online algorithm is compared with the performance defined for the optimal offline algorithm of the problem. The baseline for the concept of the optimal offline algorithm of an online problem is that this algorithm has a priori knowledge of the entire input and thus is able to confront with it with minimum cost. The metric used to describe the performance of an online algorithm A is the competitive ratio. The competitive ratio is always computed against an adversary knowing in advance the operation of the online algorithm and generating a sequence σ, based on this knowledge, with the online algorithm serving it. To compute the competitiveness of an online algorithm A, if $A(\sigma)$ stands for the cost of the online algorithm and $OPT(\sigma)$ stands for the cost of the optimal offline algorithm the online algorithm A is c-competitive if a constant α exists that $A(\sigma) - c \cdot OPT(\sigma) \leq \alpha$.

Edge computing [3] is a virtualization technology providing computation, and networking operations between devices located at the edge of the network deployment and cloud computing. The baseline con-

*Christos Bouras, bouras@cti.gr

https://dx.doi.org/10.25046/aj030519

text of Edge computing is having light-weight devices at the proximity of end users enabling the provision of custom and location-aware services. Such micro clouds can be easily integrated to Enhanced Packet Core (EPC) of mobile networks enabling on EPC new capabilities on deploying evolutionary architectures or service solutions via virtualization or not.

This paper is an extension of work originally presented in [4] focused on the application of AL-FEC error protection on NGMN multicast services where the novel MTO adaptive online algorithm aiming to efficiently apply AL-FEC protection is analyzed. In this work, we moreover concentrate on the migration of AL-FEC protection to the edge of mobile edge networking architecture as presented in [5]. We provide an analysis on such an architecture and the operations required for such integration. Furthermore, we evaluate the AL-FEC application performance on the edge revealing, apart from the benefits introduced from such a choice, the performance bounds on its deployment.

The context of this work has the following structure: In Section 2 we discuss relevant work of this paper. In Section 3 we describe the concept of Mobile Edge computing and in Section 4 we present the AL-FEC reliability control. In Section 5 we present the proposed online algorithm and we analyze its performance. In Section 6 we analyze the migration of the AL-FEC error protection on the edge and we evaluate this design performance under several perspectives. Concluding in Section 7, we discuss the advantages introduced from the proposed online error protection scheme and the AL-FEC application on the edge and we refer to some possible future directions that could follow in order to extend this work.

2 Related Work

Many active research fields of communication networks utilize online algorithms. In the work of [6] the authors cope with routing alogorothms under energy efficiency constraints and they present an online algorithm improving the throughput on multihop networks, The authors of [7] utilize online to address energy-constrained limitations on ad-hoc networks in the context of multicast routing problems. Moreover, the authors of the work [8] provide distributed online algorithms for the frequency assignment online problem. In the work presented in [9] the authors cope again with energy efficiency in conjunction with delay constraints on wireless multicast deployments presenting an algorithm for online scheduling problems. The authors of [10] investigate the decision sending a source packet, retransmiting a lost packet or sending a redundant packet with competitive analysis, providing an online algorithm capable to pick the appropriate decision. Finally, the authors of this manuscript in [11] stated the online AL-FEC policy problem referring to reduction of the efficient selection of the introduced transmission overhead in an AL-FEC protected delivery to online problems for mobile multicast networks and they provided a first attempt on this newly introduced online problem with a randomized online algorithm. Moreover, the same authors in the work [12] designed a deterministic online algorithm utilizing weight assignments and afterwards the same authors provided an enhanced adaptive version of this deterministic online algorithm in [13].

On the edge computing field, the authors of [14] provide a thorough definition of this techonology, analyzing several case studies and discussing the challenges and opportunities of edge computing. Moreover, the authors of the owrk [15] present the edge technology in the context of sensor networks, peer-to-peer networks and network virtualisation functions. Finally, in the work [16] the authors investigate the faesibility of utilizing constrained devices as edge gateways in edge arhitectures. From the results of this work they conclude that under appropriate configuration and fine-tuning, scalability and reduced overhead can be achieved.

3 Mobile Edge Computing on 5G Multicast NGMN

The increased demands on network capacity and user experience drive the next generation mobile systems. The efficient utilization of the radio spectrum over common resources is one of the most major aspects for broadcast and multicast delivery towards 5G NGMN. Apart from the development of novel technologies there is the path where multicast services will cooperate and be enhanced by several existing ecosystems, like Internet of Things (IoT) or Device-to-Device (D2D) communication. To achieve this cooperation development is required on all layers of the protocol stack, from modulation and coding to transport protocols, and applications.

Mobile Edge Computing (MEC) [3] is a technology under standardization from the ETSI MEC Industry Specification Group (ISG). Mobile Edge Computing describes a service environment with cloud-enabled capabilities located at the edge of the mobile network in proximity to mobile end-users. The main goals of MEC is to achieve reduced latency, increased efficiency for network operation and services delivery, and as a result an enhanced user experience. MEC side by side with Functions Virtualization (NFV) and Software-Defined Networking (SDN)) [17] is one of the main focuses on the development of 5G evolved systems. One of the main advantages of MEC in conjunction with NFV and SDN is that will enable the transformation of traditional deployments to a programmable interface for various operations contributing on the achievement of the demanding requirements of 5G standard. MEC is itself based on virtualization, and comes to complement NFV, while NVF focuses on network functions while, MEC is responsible to enable applications running at the edge of the network. Another advantage of MEC is the fact that allows operators to reuse their infrastructure deployments since the infrastructure

hosting MEC operations is similar with the equired NFV infrastructure hence, it is possibl to host both VNFs (Virtual Network Functions) and MEC applications on the same infrastructure. Combining all of the above, MEC technology will allow 5G environments to provide improved latency, bandwidth, and location awareness and so new or existing services/operations could be hosted on the edge.

4 AL-FEC Error Protection

Fountain Codes are codes developed to provide reliability control on data delivery over unreliable channels characterized by unspecified probability of data corruptions [18]. Fountain code name comes from the fact that operate like water fountains infinite supplying water drops and anyone collecting those drops in a bucket under the fountain with the bucket removed when enough drops are collected. In the same context but with a digital source instead of a water fountain, a recipient receives encoded symbols from one or more transmitters and collects them till enough are obtained. Then the recipient is able to reconstruct the original transmitted object, and which symbols are received does not matter. Since the source can theoretically generate unbounded amount of encoded symbols, Fountain Codes are rateless codes.

Since random linear codes cannot be applied in proactice due to their complexity, Luby Transform (LT) codes presented in [19] which are able to provide applicable encoding and decoding and at the same time they are able to maintain low transmission overhead. In more details, LT codes can operate close enough to channel encoding under appropriate selecting the distributions of the edges in the Tanner graph. Moreover, if the reliability of the decoding process is decreased the complexity can be further reduced providing linear time for encoding and decoding complexity. Since all of the input symbols cannot be decoded with the lower degree distribution for the same overhead if an erasure correcting pre-code is utilized the erasures arising from the weakened decoder could be recovered. Raptor codes provide exceptional encoding and decoding performance if a linear time block code like LDPC is used for the precoding process.

RaptorQ codes [20] are the predecessor of Raptor codes providing much more efficiency on AL-FEC encoding and decoding process. They also present by far improved flexibility and error recovery. Encoding of RaptorQ codes is almost the same with Raptor code encoding. Of course RaptorQ codes introduced an enhanced design providing enhanced performance compared with their predecessor. The main difference on the operation of RaptorQ and Raptor codes is that the newly introduced RaptorQ codes utilize larger finite fields overcoming the performance limitations of Raptor code and achieving recovery from lost symbols with decreased transmission overhead. Apart from this RaptorQ codes are capable to operate over higher number of source symbols and thus increased number of generated encoding symbols. This expansion on the number of symbols RaptorQ codes can handle enables the simplification and increased flexibility of applying AL-FEC protection.

5 Online AL-FEC Protection

5.1 Online Model

In this work we assume a common mobile networks environment where multicast delivery of streaming flows to endusers is simulated. In more details, a multicast source injecting bunches of packets in the network which are handled as a continuous object and are delivered to endusers via multicast bearers encapsulated in RTP/UDP streaming flows.

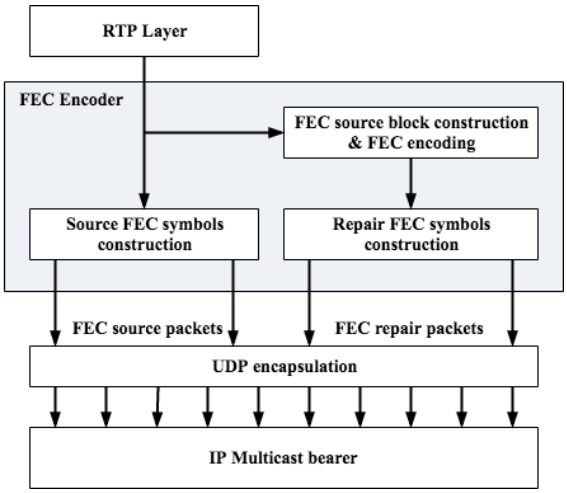


Figure 1: AL-FEC encoding process

Regarding the application of AL-FEC protection in the transmitted object we utilize the newly introduced RaptorQ FEC codes. Fig. 1 presents the encoding process of AL-FEC protection The first step is to partition the transmitted object into one or more FEC source blocks with each one of the FEC source blocks having size k symbols, a value which is an encoding parameter and is called source block length (sbl). During the RaptorQ AL-FEC encoding process a number of redundact FEC symbols are generated from each FEC source block. Those redundant symbols are called repair symbols. The amount of generated repair FEC symbols depends on the desired AL-FEC protection introduced by the multicast source. The outcome of the RaptorQ AL-FEC encoding is a certain amount of FEC encoding symbols which can be either source or repair symbols. In the simulated environment each produced FEC encoding symbol receives a number which is unique for each one.

The decoding failure probability which is the metric characterizing the decoding performance of RaptorQ AL-FEC is described as (1) [21]:

$$p_{f_{RQ}}(n,k) = \begin{cases} 1, & \text{if } n < k \\ 0.01 \times 0.01^{n-k}, & \text{if } n \geq k \end{cases} \quad (1)$$

For a FEC source block of size k with an enduser received n encoding symbols, the quantity $p_{f_{RQ}}(n,k)$ is the probability the enduser to fail to decode this source block. Regarding the successful reconstruction of a transmitted FEC source block From the receiver perspective with respect to the previous modeled AL-FEC decoding failure probability we assume, forced from the results available in [22], that the threshold for the decoding failure probability of each FEC source block can be 10^{-2} to mark the FEC source block as successfully reconstructed or not.

Finally, regarding the multicast bearer level we model the delivery or not of a transmitted packet with an independent loss transcript. That means that on each packet sequence transmitted on each enduser an independent packet loss mask is applied according to the desired packet loss rate defined in the environment. In this work p denotes the average network packet loss rate of the mobile multicast bearers.

5.2 Optimal Offline Algorithm

To provide the competitive analysis of the proposed online algorithm we need to utilize the optimal offline algorithm of the online problem. Since the optimal offline algorithm has a priori knowledge of the problem input, and according to the work presented in [23], the optimal offline algorithm in the context of a multicast source which has to select the appropriate amount of transmission overhead introduced in a delivery is a multicast source which will select the introduced transmission overhead to a value close to the average packet loss rate of the network.

Moreover, based on the enhanced recovery performance properties of RaptorQ codes as described in (1) and based on the assumption of this work for the failure probability of the decoding process of a received FEC protected source block the optimal offline algorithm can achieve the specific threshold introducing no more FEC repair symbols than the average number of lost encoding symbols computed by the average packet loss rate p of the multicast environment.

According to previous analysis, given that sbl is the number of symbols a source block contains and r is the number of repair symbols the optimal offline algorithm decides to introduce in each FEC source block r can be calculated as: $(sbl + r) \cdot p$. Hence, the cost OPT of the optimal offline algorithm is: $OPT = sbl + r$.

5.3 Mean Transmission Overhead Algorithm

The baseline of the Mean Transmission Overhead (MTO) algorithm is the adaptation of the AL-FEC overhead based on feedback received from UEs of past deliveries. Hence, the operation of the MTO algorithm depends on the previous transmission rounds. On each round the input of the algorithm is the UE coverage attribute. UE coverage denotes a threshold on the number of UEs participating on the delivery and is used to indicate if the transmission round should be considered as sufficient protected or not. The outcome of the MTO algorithm is the computation of the mean AL-FEC overhead after each round and the update of the introduced transmission overhead with the computed value for the current round.

Each UE is able to report the outcome of the AL-FEC decoding process on the transmitted object enabling the algorithm to compute the overall UE coverage of a completed transmission round. Based on this, the MTO algorithm is able to compute the attribute ideal transmission overhead for each past transmission round. This attribute denotes the transmission overhead which should had been introduced in the delivery round to achieve the defined UE coverage for this round. After computing this ideal value transmission overhead of the completed round the algorithm is ready to update the value of the mean transmission with it. Finally, the MTO algorithm will update the current transmission overhead that will be introduced in the current transmission round with the updated value of the mean transmission overhead.

At this point, clarifying the feasibility of a crucial part on the MTO operation, i.e., the feedback received on the reception status of the transmission object from each UE, we have to mention that a various of mobile multicast standards describe a post-delivery process where the UE is able to provide feedback to the multicast source on several aspects i.e., report on delivery status, file repair status, etc. Such kind of capabilities enable a multicast source to track the status of a transmission and maintain extensive report information for past content deliveries.

Having analyzed the operation of the MTO algorithm we can compute the cost of the online algorithm. Denoting with r' the number of source symbols that the online algorithm will introduce to the transmission of a FEC source block with length sbl the value of r' is computed according to the attribute mean transmission overhead of the algorithm and the cost of the online MTO algorithm is: $ALG = sbl + r'$.

5.4 Performance Evaluation

Block Length First part of performance evaluation introduce a comparison on the percentage of UEs that are satisfied. Satisfied UEs are the ones which are able to reconstruct the original object. The comparison is done between the performance of the MTO online algorithm, an approach where AL-FEC overhead is fixed to a specific value approach performance, and lastly a common error recovery approach which is feedback-based and lost packets are retransmitted to individual UEs.

Fig. 2 contains results for a transmitted entity encoded to source blocks with the size of each block varying between 2048 to 65536 symbols. The evaluated packet loss rate is 5%. 100 participating UEs in the multicast delivery and a 65536 symbol sized transmitted object are the other parameters utilized in this evaluation. Moreover, 90% is the target coverage for

Algorithm 1 Mean Transmission Overhead

```
1: procedure (ue_coverage)
2:     current_ideal_overhead ← computeIdealOverhead(ue_coverage)
3:     mean_overhead ← updateMeanOverhead(current_ideal_overhead)
4:     transmission_overhead ← mean_overhead
5: end procedure
```

this evaluation.

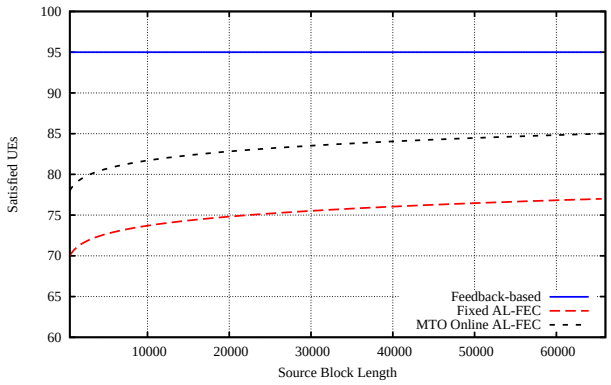


Figure 2: Satisfied UEs vs. Source Block Length

As was anticipated, we cannot have results regarding *sbl* increase for feedback-based packets retransmission case since AL-FEC protection is not used there at all. We can just note that the feedback-based approach is capable to maintain continuous and high volumes of satisfied UEs. However, this performance comes at its own expense, a fact discussed in next subsections of simulation results. Discussing the performance of the approaches utilizing AL-FEC, the performance enhancement by increasing the length of the source block, is clearly noticeable. RaptorQ codes are able to operate more efficiently by increasing the *sbl* since they are able to exploit protection spreading across the whole encoded entity.

In addition, the curve trend for the quantity of the satisfied UEs we notice that is exact same for the MTO algorithm, as well as for the fixed AL-FEC transmission overhead policy, since the evaluated loss rate remains fixed, hence packet delivery conditions are not altered for each simulation snapshot and the MTO algorithm operation is founded on the adaptation of the transmission overhead according to previous reception conditions.

Packet Loss Rate In this section, the network's packet loss rate impact on the performance of the MTO algorithm is analyzed. In Fig.3 there is a presentation of the achieved return on satisfied UEs, for different packet loss rates relative to the approach introducing a constant amount of overhead and the approach where lost data are retransmitted on UEs request. On this evalutation the packet loss rate is gradually increased from 1% to 20% while delivering a 4096 packet length object to 100 UEs, assuming 4096 symbol sized AL-FEC source blocks and one packet per symbol.

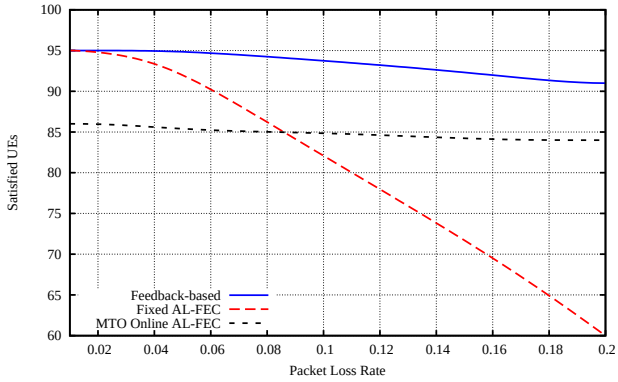


Figure 3: Satisfied UEs vs. Packet Loss Rate

Once more, we can point out that the efficiency of the packet retransmission is superior to the one of the fixed transmission overhead protection. An expected remark is that, for only low average packet loss rates, the efficiency achieved is identical to the constant overhead protection. This is based on the fact that the introduced AL-FEC overhead is of fixed value for the fixed policy approach chosen at a higher value in comparison to the packet loss rate, and causes wasting valuable network resources. Another remark regarding the efficiency of the lost packets retransmission protection is the fact that, even slightly, it is reduced with the packet losses increase. This is an expected behavior, as for high rates of packet losses, the amount of UEs that will enter in bad reception or even dropped will constantly increase.

As shown in the curve trend, even though the performance would be ephemeral, the constant overhead approach can achieve high volumes of satisfied UEs for low average packet loss rates but this is observed just initially, i.e., for packet losses up to 2%. More on this, we can indicate that the constant overhead performance declines fast and the UEs coverage becomes 25% less than the MTO algorithm performance for packet loss 20% since the constant overhead scheme cannot respond to the increase of the packet loss. Regarding the MTO algorithm we observe that now the performance is constant. Of course, the coverage achieved is less than the lost data retransmission approach but, this has nothing to do with the MTO performance and it is just a consequence of the UEs coverage target of the simulation.

Network Resources Those latter simulations of Fig.4, examine the performance of the three evaluated approaches on a network utilization perspective i.e. the amount of exchanged packets during a multi-

cast delivery, against the increasing simulated average packet loss rates is examined. Once more, the packet loss rate ranges from 1% to 20%. For this evaluation we perform a simulation of a 4096 packet object transmission to 100 UEs. As for the AL-FEC protection parameters, we use a 4096 symbol sized source block. It is important to clarify that for the feedback-based protection scheme each UE is able to provide feedback, about the packets that need to be retransmitted, to the multicast source, through unicast bearers. Then, in a following repair transmission phase, the packets will be transmitted to the appropriate UEs from the multicast source, through unicast bearers. Finally, for AL-FEC protection application we assume exactly one FEC symbol per packet.

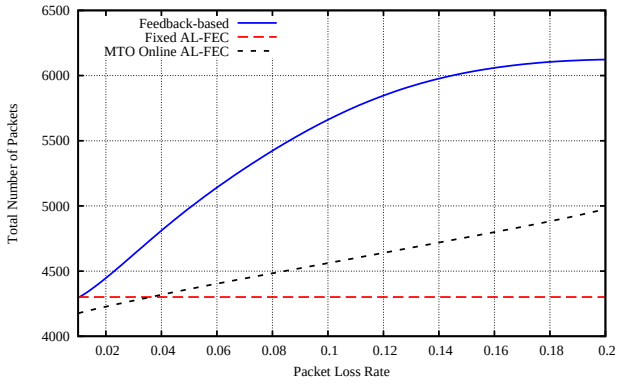


Figure 4: Total Number of Packets vs. Packet Loss Rate

The first immediate remark comes for the constant overhead approach where the amount of packets exchanged remains constant too for any evaluated packet losses value. Of course this is the expected performance for this approach as previously analyzed. The cost regarding network resources to achieve the high performance noted for the retransmission based approach as discussed in previous parts is revealed in this subsection. Compared with MTO, This approach requires a significantly higher number of packets transmitted across the network, and more specifically, we can notice that for packet losses higher than 10% this approach adds on the network over 50% of additional traffic. Finally, on the MTO approach, the curve for the number of exchanged traffic matches the network packet losses trend. This is a direct consequence of the algorithm's nature to adapt the overhead to the reception conditions. Although MTO's approach can not reach the error recovery provided by the retransmission approach as discussed in previous part, its profit on resource utilization is higher than the degradation on error protection.

6 AL-FEC Error Control on the Edge

In this second part we further analyze the AL-FEC Protection Scheme integrated with MEC presented in [5]. We provide performance evaluation results discussing initially the improvements the proposed scheme could introduce in multicast NGMN but apart from this we reveal the bounds on the performance of this scheme regarding deployment parameters.

Fig. 5 presents the basis of the architecture where the AL-FEC protection is migrated to the edge gateways of the network. While in conventional deployments the AL-FEC is applied on the multicast gateway level, in this work the AL-FEC application is also handled by the edge gateways responsible for a cluster of UEs in cooperation with the multicast gateway.

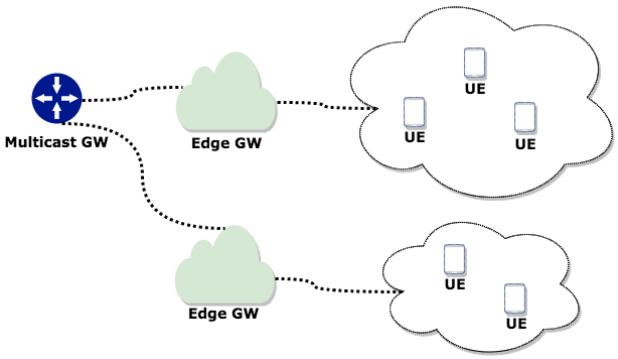


Figure 5: AL-FEC error control on the edge

The delegation of the decision on the transmission overhead amount to the level of the edge gateways enables a coarse-grained policy on the process computing the most appropriate AL-FEC transmission overhead for a delivery. Hence, this architecture provides enhanced efficiency for AL-FEC since the number of repair packets will be reduced resulting also on network load decrease always with respect to the reception conditions of the UEs cluster. Another important fact that will also provide enhanced scalability and efficiency is that any monitoring functions required for an AL-FEC scheme can also be implemented on the edge gateways.

In more details, the Multicast GW is responsible to select, according to the Multicast GW and Edge GW connection type, if it will introduce AL-FEC redundancy to the content forwarded to the edge layer. On the edge GW layer when the transmitted content arrives the GW should decide if the content will be forwarded as it is to the endusers in case of an already encoded content. In case of the arrival of a non AL-FEC protected content, the edge GW will AL-FEC encode it introducing the appropriate overhead based on networking functions implementation on the receptions conditions of the UEs cluster served by the specific Edge GW.

For the sake of clarity, we have to clarify that in the application layer, apart from the AL-FEC protection applied, a file-repair post-session is also utilized [24] where a UE that fails to decode a FEC protected source block is able to determine which source symbols are missing for the source block and achieve the retransmission of them via dedicated unicast bearers.

6.1 Performance Evaluation

In this part we present theoretical performance results on the potential improvements such a scheme has on the transmitted redundancy and in general in the overall network traffic.

AL-FEC Transmission Overhead In the presented results we investigate the improvement on the volume of the data redundancy introduced. The percentage results refer to the comparison of the proposed method with conventional architectures with the protection redundancy computed from the multicast source. In both cases we assume that both examined approaches require the same delivery success rate in the endusers.

In Fig. 6 we investigate the decrease on the average introduced error protection overhead with the increase of the number of UEs per each Edge GW for three different values of average packet loss rate 5%, 10% and 20%.

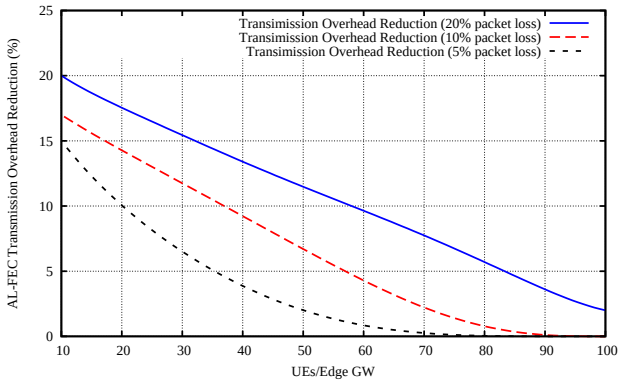


Figure 6: AL-FEC Transmission Overhead vs UEs/Edge GW

It is anticipated that as the number of UEs served by an Edge GW increase the gains on the transmission overhead are decreased since the required average transmission overhead is increased. In more details when evaluating 10 UEs per Edge GW with 20% packets lost the transmission overhead is reduced around 20% and for the same number of UEs but for 10% and 5% of average packet loss rate the improvement is about 17% and 15% respectively. For the case of 20% of average packet loss rate and for 50 UEs per Edge GW we notice a decrease of almost 10%. While, for 100 UEs the decrease is shortened to around 5%. This performance is expected since if the number of endusers assigned to an Edge GW increases then the number of UEs with heterogeneous packet losses under the same Edge GW is increased too resulting to the increase of the average transmission overhead amount that should be introduced to the delivery. However, the improvement on the introduced AL-FEC transmission overhead is still noticeable even in the case of the 100 UEs per Edge GW and 20% packet losses. Another interesting remark is that for the cases of evaluated packet losses of 5% and 10% a threshold for the efficiency of the proposed scheme is revealed with respect to the number of UEs served by an Edge GW since for 5% packet losses we have no gains for more than around 50 UEs and for 10% packet losses this threshold is less than 70 UEs.

Network Traffic In this last part we investigate the impacts of the edge error control architecture on the network traffic of the mobile network. We evaluate the total amount of traffic towards the UEs but also on the backwards channels too since a UE is able to request the retransmission of lost symbols from the multicast source via unicast bearers in case of corrupted source content.

Fig. 7 presents simulation results evaluating the impact on the network traffic in terms of packets exchanged on the network with the average packet loss rate increase from 5% up to 20% and the number of UEs served by an Edge GW increased from 10 UEs to 100 UEs.

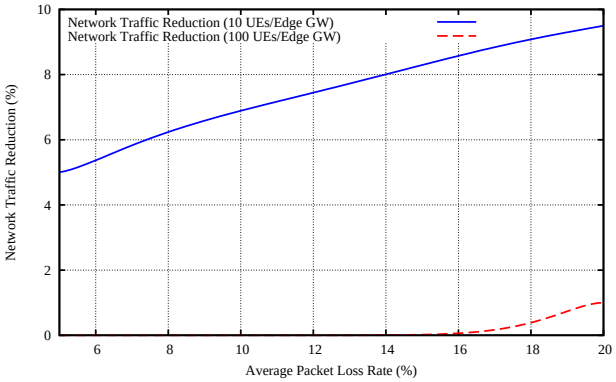


Figure 7: Network Traffic vs Packet Loss Rate

The first remark from the trends of the 10 UEs curve is that the network traffic decrease is enhanced as the packet losses increase. For 10 UEs served by an Edge GW with 5% of packet losses the network traffic decrease is more than 5% and for 20% of losses the decrease overcomes 9%. This is a direct consequence from the results presented in previous section since as the packet losses are increased the decrease achieved from the proposed scheme on the average transmission overhead required is increased too. Hence, less redundant repair symbols are transmitted required in average to achieve a successful decoding. On the other hand, as in the previous results presented in Fig. 6 the increase of the UEs delegated to an Edge GW to 100 UEs reveals again the scalability bound of the proposed scheme since we observe that in this case the decrease on the network case is eliminated for average packet loss rate values till around 18% and only extremely high packet losses of 20% denote a network traffic decrease of around 1%. This is a direct consequence of the fact that as the UEs delegated to an Edge GW increase the coarse-grained tuning on the required AL-FEC transmission overhead gets eliminated since the number of UEs with high deviations on reception conditions is constantly increased too.

7 Conclusions and Future Work

The MTO online algorithm is presented in this manuscript on the online AL-FEC policy problem. The baseline of the proposed algorithm is the adaptation of the introduced AL-FEC transmission redundancy based on the feedback received from previous transmissions. We have presented and stated the introduced MTO AL-FEC online algorithm based on the online AL-FEC policy and we have provided an analysis of the impacts of this algorithm on the application of AL-FEC error control comparing with feedback-based error recovery and the common approach of introducing a fixed amount of AL-FEC overhead in the transmission and providing simulation results on the achieved performance of the MTO online algorithm. According to the results, the online algorithm achieves close performance to that of the retransmission based approach and sometimes it overcomes it. Another aspect we have investigated is the impacts of the MTO algorithm on network resource utilization. The evaluated results verified that the online approach is the most efficient compared with the other evaluated approaches.

The second part investigates the migration of AL-FEC protection on the edge of a MEC-enabled network. Apart from this, performance evaluation results were presented indicating the gains that such a scheme can introduce on the efficiency of the AL-FEC protection application since, such a choice is able to introduce coarse-grained error recovery capabilities on a multicast network. This design choice was analyzed in terms of network traffic reduction but also we have investigated the scalability bounds of the proposed scheme under different reception conditions and deployment parameters. We have verified that, under the condition that the parameters of such an approach are carefully selected, such a design choice id capable to reduce the AL-FEC overhead and hence increase network resources utilization.

Future work of this manuscript could concentrate on the design of more sophisticated and advanced online algorithms on the problem investigated in this work i.e., advanced approaches to apply AL-FEC efficiently on NGMN. Another potential field that could extend the work presented in this manuscript is device-to-device (D2D) communication over MEC architectures. IoT architectures, apart from the common communication channels providing connectivity between the cloud and IoT devices, defines also connectivity capabilities between the IoT devices. Based on this, the proposed architecture could be enhanced enabling D2D capabilities where, during the AL-FEC source block reconstruction, missing repair symbols could be also received from another device in proximity and not from the Edge GW.

References

[1] A. Borodin and R. El-Yaniv, *Online computation and competitive analysis*. New York, NY, USA: Cambridge University Press, 1998.

[2] R. Motwani and P. Raghavan, "Algorithms and theory of computation handbook," M. J. Atallah and M. Blanton, Eds. Chapman & Hall/CRC, 2010, ch. Randomized algorithms. [Online]. Available: http://dl.acm.org/citation.cfm?id=1882757.1882769

[3] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing?a key technology towards 5g."

[4] C. Bouras and N. Kanakis, "Evolving al-fec application towards 5g ngmn," in *2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Feb 2018, pp. 1–5.

[5] C. Bouras and N. Kanakis, "Al-fec application on ngmn-edge computing integrated systems," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2018, pp. 364–369.

[6] L. Lin, N. Shroff, and R. Srikant, "Asymptotically optimal energy-aware routing for multihop wireless networks with renewable energy sources," *Networking, IEEE/ACM Transactions on*, vol. 15, no. 5, pp. 1021 –1034, oct. 2007.

[7] W. Liang and X. Quo, "Online multicasting for network capacity maximization in energy-constrained ad hoc networks," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 9, pp. 1215 –1227, sept. 2006.

[8] J. Janssen, D. Krizanc, L. Narayanan, and S. Shende, "Distributed Online Frequency Assignment in Cellular Networks," *Journal of Algorithms*, vol. 36, no. 2, pp. 119 – 151, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0196677499910684

[9] A. El Gamal, C. Nair, B. Prabhakar, E. Uysal-Biyikoglu, and S. Zahedi, "Energy-efficient scheduling of packet transmissions over wireless networks," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3, 2002, pp. 1773 – 1782 vol.3.

[10] Y. Bartal, J. Byers, M. Luby, and D. Raz, "Feedback-free multicast prefix protocols," in *Computers and Communications, 1998. ISCC '98. Proceedings. Third IEEE Symposium on*, jun-2 jul 1998, pp. 135–141.

[11] C. Bouras, N. Kanakis, V. Kokkinos, and A. Papazois, "Deploying AL-FEC with Online Algorithms," in *Next Generation Mobile Apps, Services and Technologies (NGMAST), 2013 Seventh International Conference on*. IEEE, 2013, pp. 175–180.

[12] C. Bouras and N. Kanakis, "A Competitive AL-FEC Framework over Mobile Multicast Delivery," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*, jul. 2013.

[13] C. Bouras and N. Kanakis, "An Adaptive Weighted Online AL-FEC Algorithm over Mobile Multicast Networks," in *Wireless Communications and Networking Conference (WCNC), 2014 IEEE*. IEEE, 2014.

[14] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, Oct 2016.

[15] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 27–32, Oct. 2014. [Online]. Available: http://doi.acm.org/10.1145/2677046.2677052

[16] P. Bellavista and A. Zanni, "Feasibility of fog computing deployment based on docker containerization over raspberrypi," in *Proceedings of the 18th International Conference on Distributed Computing and Networking*, ser. ICDCN '17. New York, NY, USA: ACM, 2017, pp. 16:1–16:10. [Online]. Available: http://doi.acm.org/10.1145/3007748.3007777

[17] 5G Infrastructure Public Private Partnership, "5G Vision: The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services," 5G Infrastructure Public Private Partnership, https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf, Tech. Rep., 2015.

[18] D. J. MacKay, "Fountain codes," in *Communications, IEE Proceedings-*, vol. 152, no. 6. IET, 2005, pp. 1062–1068.

[19] M. Luby, "LT codes," in *Foundations of Computer Science, 2002. Proceedings. The 43rd Annual IEEE Symposium on*, 2002, pp. 271–280.

[20] M. Luby, A. Shokrollahi, M. Watson, T. Stockhammer, and L. Minder, "RaptorQ Forward Error Correction Scheme for Object Delivery," RFC 6330, Internet Engineering Task Force, Aug. 2011. [Online]. Available: http://tools.ietf.org/rfc/rfc6330.txt

[21] 3GPP, "Rationale for MBMS AL-FEC Enhancements," 3rd Generation Partnership Project (3GPP), Tdoc S4-110449, 2011.

[22] 3GPP, "Simulation results for the performance and complexity of RS codes for MBMS FEC," 3rd Generation Partnership Project (3GPP), Tdoc S4-050107, 2005.

[23] C. Bouras, N. Kanakis, V. Kokkinos, and A. Papazois, "Application layer forward error correction for multicast streaming over LTE networks," *International Journal of Communication Systems*, 2012. [Online]. Available: http://dx.doi.org/10.1002/dac.2321

[24] 3GPP, "Multimedia Broadcast/Multicast Service (MBMS); Protocols and codecs (Release 10)," 3rd Generation Partnership Project (3GPP), TS 26.346, 2011. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/26346.htm

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 3, 161-167 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Fuzzy Logic Based Selective Harmonic Elimination for Single Phase Inverters

Zeynep Bala Duranay *, Hanifi Guldemir

Firat University, Technology Faculty, Electrical-Electronics Engineering, 23119, Elazig, Turkey

ARTICLE INFO

Article history:
Received: 04 May, 2018
Accepted: 22 June, 2018
Online: 26 June, 2018

Keywords:
Fuzzy logic
Harmonic analysis
Harmonic elimination
Inverter
Pulse width modulation

ABSTRACT

A selective harmonic elimination system using fuzzy logic for the elimination of high magnitude harmonics with frequencies close to fundamental in the output voltage of single phase inverters is presented. The system does not require look up tables for storage of the data as in traditional harmonic elimination methods. The input of the fuzzy system is the modulation index values. The output of the fuzzy system provides the switching angles which are further used to construct the switching signal for the switches in the inverter. With this fuzzy logic based selective harmonic elimination system, predetermined dominant low rank harmonics are successfully eliminated. Simulations are made with Matlab/Simulink and the results are presented which show the effectiveness of the presented harmonic elimination system.

1. Introduction

This paper is an extension of work originally presented in 2017 XXVI International Scientific Conference Electronics (ET) [1]. The use of power electronic devices in industrial and consumer applications has resulted nonlinear sinusoidal voltage and current to be drawn from the source. These nonlinear loads distort the sinusoidal form of the alternating current which results harmonics in the electrical system. The power quality of an electrical system is determined by its harmonic content. These harmonics may be classified as voltage or current harmonics. These harmonics can occur either supply or load. Current harmonics are generated by the harmonics of the source voltage which are depend on the type of the load connected to supply which can be resistive, inductive or capacitive. Nonlinear operation of power converters feeding loads cause harmonics to be produced in the load. These harmonics cause extra heating in transformers and motors.

Supply harmonics are produced by the source with nonsinusoidal voltage or current waveforms. Supply harmonics produce extra loss, cause electromagnetic interference and ripple torques in ac motors [2].

A pulse width modulated inverter which converts dc source to ac supply with desired voltage and frequency is used in many electrical devices such as uninterruptable power supplies (ups), switch mode power supplies(smps), machine drives. However, it is seen that the output voltage of the inverters has higher harmonic content.

Among others [3-4], the pulse width modulation (pwm) technique is one of the commonly applied modulation strategy which controls the amplitude and frequency of the voltage produced by inverters.

Pwm techniques have been widely used in converter and inverter control where the switches are switched on and off number of times in each period. The controlled output signal is obtained by changing the width of these pulses.

Different pwm methods have been developed to suppress harmonics in converters and inverters. These techniques can be summarized as

- Carrieer based pwm
- Space vector pwm
- Third harmonic injection pwm
- Selective harmonic elimination pwm

One of the effective pwm method for the suppression of harmonics from the spectrum of the inverter output voltage is the selective harmonic elimination (SHE) technique [5-7].

The voltage or current waveform is represented by Fourier series expansion in SHE technique. The coefficients of the harmonics those will be eliminated are set to zero and the

*Zeynep Bala Duranay, Firat University Electrical-Electronics Engineering 23119 Elazig Turkey, zbduranay@firat.edu.tr

coefficient of the fundamental is set to the desired value. Hence, a set of nonlinear equation is obtained. This equation set need to be solved in terms of unknown switching angles for each value of modulation index. The number of equation depends on the number of harmonics to be eliminated [1].

For the offline approach the nonlinear equation set is solved for every modulation index values and stored in look up tables. These modulation index values used in the microcomputer for generating switching instants. A huge number of tables of data are required to be saved which also require large memory. This problem can be avoided by online approach. Therefore, intelligent methods are proposed for the SHE.

In this study, a fuzzy logic (FL) based SHE is presented. FL is used to obtain switching moments for the switching elements in the single-phase inverter.

The main contribution using FL is the fast response in real time operation constructing the pwm patterns for improving the inverter output voltage waveform. Whereas, in conventional methods the calculations of the switching instants are made off line. The controller needs a look up table to store the switching instants for SHE.

2. Selective Harmonic Elimination

SHE is a modulation method aiming to obtain the appropriate switching instants to suppress the low order harmonics. In SHE the switching moments are determined by using the desired magnitude of the fundamental and harmonics to be suppressed. Main aim is to determine switching angles so that the fundamental component is at the desired value and undesired harmonics which are the lower order harmonics are eliminated. Thus, Total Harmonic Distortion (THD) of the output voltage is also minimized.

Figure 1 Shows a single phase full bridge inverter structure. In conventional pwm inverters the gate signals for the switching elements in the first leg are produced by comparing a sinusoidal modulation signal with a triangular carrier wave.

180° phase shifted signals of the first leg are applied to the second leg of the inverter. Thus, the gate signals for the switches S1, S4 and S2, S3 are synchronized.

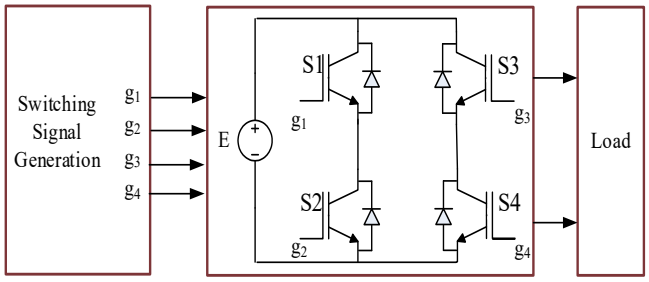


Figure 1. Single-phase symmetric pwm inverter

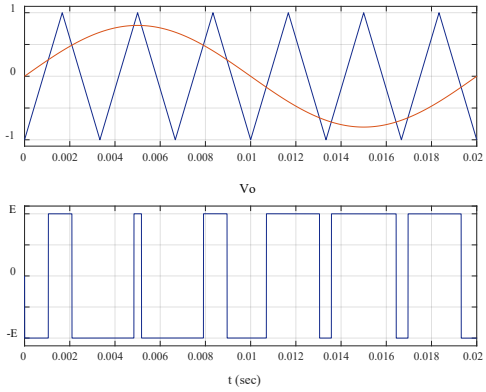
The technique used to eliminate or minimize low rank harmonics while keeping the magnitude of the fundamental component at desired value by appropriately choosing the switching angles of the voltage waveform of an inverter is known as the SHE [8-11]. With this technique, the magnitude of the fundamental component and the harmonics to be eliminated are determined and the corresponding switching angles are calculated.

The output voltage waveshape may be either bipolar or unipolar as in Figure 2 which shows normalized output voltage waveforms. In bipolar waveform, the output voltage is two level in which both +E and –E voltage pulses are used in each half period. In unipolar waveform, the output voltage is three level and only one of +E or –E voltage pulses are used in each half period. The output voltage waveshape of the SHE modulated inverter is constructed to have a quarter wave symmetry.

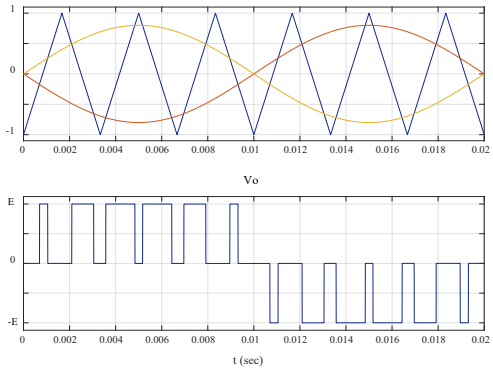
The two switches are not turned on or off in the same inverter leg at the same time. They are operated in a complementary manner that is, one switch turned on and other turned off in the same inverter leg. Thus, inverter need only two switching signals for the gate of upper switches which are produced by comparing a sinusoidal modulating wave and triangular carrier wave.

In the unipolar modulation two sinusoidal modulating wave with same magnitude but 180° out of phase are used. Switching signals for the gates of upper switches in the inverter legs are produced by comparing these two modulating waves with a sawtooth carrier.

The unipolar voltage waveform contains three states as presented in Figure 3. Signal pulses are generated with switching instants (α_1, α_2, α_3,..., α_N).



(a)



(b)

Figure 2. Bipolar and unipolar waveforms

Periodic signals can be constructed from a sum of sine and cosine functions. by a fundamental and a set of harmonic components. The coefficients of these functions can be obtained by applying Fourier transformation. Thus, the Fourier representation of the inverter output voltage is;

$$V_0 = a_0 + \sum_{n=1}^{\infty} A_n \cos(n\omega t) + \sum_{n=1}^{\infty} B_n \sin(n\omega t) \quad (1)$$

where

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} V_0(t)dt \quad (2)$$

$$A_n = \frac{4E}{\pi n}[-\sum_{i=1}^{N}(-1)^n \sin n\alpha_i] \quad (3)$$

Due to odd and quarter wave symmetry, even harmonics do not exist in the output voltage waveform and Fourier coefficients An and a_0 are equal to zero [12].

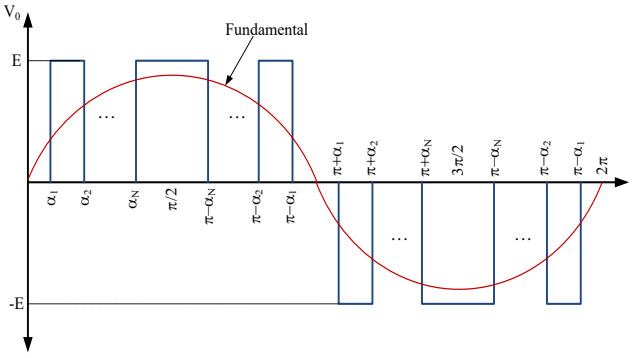


Figure 3. Inverter output voltage

Only the odd harmonics with sine components exist in the output voltage, thus (1) can be written as:

$$V_0 = \sum_{n=1}^{\infty} B_n \sin(n\omega t) \quad n = 1, 3, 5, \ldots, \infty) \quad (4)$$

where Bn is the amplitude of the nth harmonic voltage and is given by

$$B_n = \frac{4E}{\pi n}\sum_{i=1}^{N}[-(-1)^n \cos(n\alpha_i)] \quad (5)$$

and

$$0 < \alpha_1 < \alpha_2 < \cdots < \alpha_N < \frac{\pi}{2} \quad (6)$$

E is the dc bus voltage, $\omega=2\pi f_1$, f_1 is the frequency of the fundamental component and *N* is the number of switching angles per quarter cycle.

The goal of the SHE method is to determine the switching angles $\alpha_1, \alpha_2, \ldots, \alpha_N$ in the inverter output voltage so that the amplitudes of the harmonic components $B_3, B_5, \ldots, B_{N-1}$ are zero and the fundamental component B_1 equal to the desired amplitude.

Applying the constraints defined above and using (5), the following equation set can be obtained

$$\begin{bmatrix} B_1 \\ B_3 \\ B_5 \\ \vdots \\ B_n \end{bmatrix} = \begin{bmatrix} \cos(\alpha_1) & -\cos(\alpha_2) & \cdots & \pm\cos(\alpha_n) \\ \cos(3\alpha_1) & -\cos(3\alpha_2) & \cdots & \pm\cos(3\alpha_n) \\ \cos(5\alpha_1) & -\cos(5\alpha_2) & \cdots & \pm\cos(5\alpha_n) \\ \vdots & \vdots & & \vdots \\ \cos(n\alpha_1) & -\cos(n\alpha_2) & \cdots & \pm\cos(n\alpha_n) \end{bmatrix} = \begin{bmatrix} 4E/\pi \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (7)$$

where $n=1, 3, 5 \ldots N$ and *N* is the number of pulse in every quarter cycle.

The first equation in this equation set is used to adjust the amplitude of the fundamental and the others are used for the suppression of harmonics.

Thus, *N-1* harmonics can be eliminated by calculating *N* switching angles [13, 14].

If the amplitude of the fundamental component B_1 need to be adjusted then

$$B_1 = m\frac{4E}{\pi} \quad (8)$$

where m is the modulation index which is the ratio of fundamental to the maximum obtainable fundamental.

The set of nonlinear equations in (7) need to be solved using numerical methods [15, 16] to obtain the switching instants.

In this study, Newton-Raphson (NR) method, as in [17, 18] which is developed for the solution of nonlinear equations is used to solve the nonlinear equation set given in (7). The switching angles are obtained by a program developed in Matlab for solving the nonlinear algebraic transcendental equations. Solution obtained from NR is used to setup the input/output membership functions in the fuzzy controller. Table 1 lists the angles for different m values for 10 harmonics to be eliminated.

Table 1. Switching angles corresponding to modulation index

m	0.1	0.2	0.4	0.6	0.8	1
α_1	14.793	14.578	14.109	13.587	12.997	12.093
α_2	15.181	15.352	15.636	15.821	15.854	15.296
α_3	29.607	29.195	28.298	27.295	26.138	24.287
α_4	30.357	30.691	31.261	31.655	31.749	30.556
α_5	44.450	43.876	42.633	41.232	39.571	36.681
α_6	45.511	45.996	46.853	47.500	47.732	45.735
α_7	59.335	58.645	57.170	55.509	53.478	49.375
α_8	60.635	61.246	62.379	63.337	63.881	60.767
α_9	74.268	73.518	71.947	70.215	68.082	62.457
α_{10}	75.718	76.422	77.791	79.092	80.213	75.560
α_{11}	89.249	88.495	86.967	85.372	83.569	75.996

3. Fuzzy Logic Controller

FL is a theory used to represent uncertain information. It is first formulated by Lotfi Zadeh in 1960s [19]. FL is a crisp model of a system. It uses rule base to convert crisp inputs to corresponding crisp output values. This technique has been successful especially when rules cannot be perfectly known and can be used in situations where the device cannot be described by mathematical models. Decision making capability of fuzzy logic controller (FLC) is achieved by human behavior.

Fuzzy systems provide a quick and approximate solution for complex systems which do not have exact mathematical model [20]. The real life and industrial applications of FL can be found in [21-23].

FL is realized in four steps which are the determination of fuzzy sets, definition of membership functions, construction of rules and defuzzification as shown in Figure 4.

Fuzzification is the first step in FL system. In this step, crisp input values are measured and converted to suitable linguistic variables using expert's knowledge and experience.

The second step is the development of rule data base which defines the principles of the controller in terms of the relationship between input and output. This step is known as Inference Engine and supplies the required information for control rules

The last step is the conversion of the fuzzy outputs to control action which is called as Defuzzification which provides an actual control action.

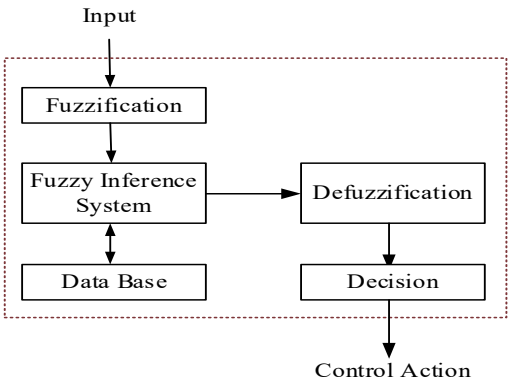


Figure 4. Block diagram of the FL system

Two types of fuzzy inference system which are Mamdami-type and Sugeno-type can be used. These types are differ according to the determination of outputs. Sugeno-type inference produces an output that is either constant or linear mathematical expression.

Due to the use of numerical values at the input and outputs, Sugeno-type fuzzy system is used for the implementation of the input and output data. Figure 5 shows the input membership function.

The outputs of the fuzzy control system which are the switching angles are computed for a given input which is the modulation index as shown in Figure 6.

The input data are represented by 10 membership functions. Because 10 harmonics are going to be eliminated which requires 11 switching angle thus 11 outputs each has 50 constant type membership functions for the representation of output data are used in the fuzzy system. Totally 104 rules are used between the input membership functions and output membership functions.

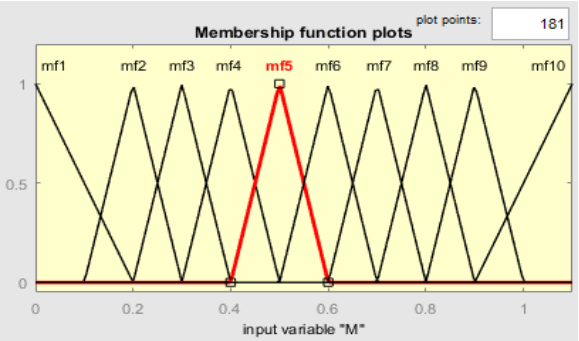


Figure 5. Input membership function

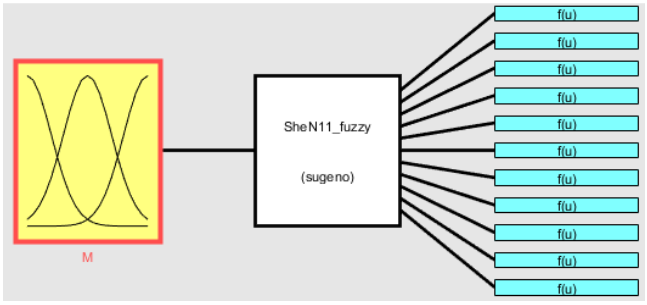


Figure 6. Fuzzy system

4. Switching Signal Generation

The switching instants corresponding to each m are determined by solving the nonlinear equation set for every value of m. These m and switching angles values are used in FLC. The m is used as the input to FLC and the output of this controller is the corresponding switching angles.

The switching instants obtained from FLC for a given m value are used as inputs for generation of switching signals.

As in traditional pwm methods, the switching angles produced by FL are compared with a timer content with specified frequency to produce pwm switching signals.

Figure 7 shows the generation of the switching signals for elimination of three harmonics i.e. N=4. In this system the switching angles are $\alpha1$=23.56º, $\alpha2$=39.26º, $\alpha3$=48.96º and $\alpha4$=89.20º. If the frequency is 50 Hz, the switching signals are obtained by comparing the timer content which is the triangular wave with the switching angles. Because of the symmetry, only the half period is given in this figure.

The Simulink block used to obtain the switching signals is shown in Figure 8.

The signals are constructed by comparison of switching instants with the sawtooth carrier and given to inverter switches as in Figure 9. Figure 10 represents a switching signals generation of 3 harmonic elimination.

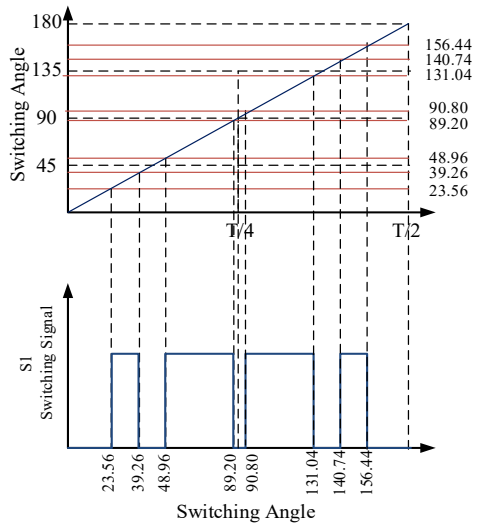


Figure 7. Switching signal generation for elimination of three harmonics

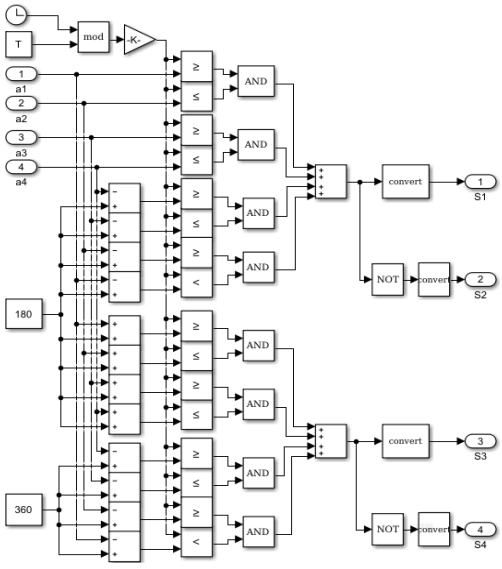


Figure 8. Simulink block for switching signal generation for the elimination of three harmonics

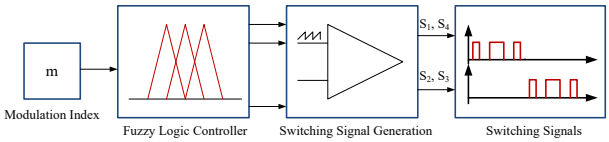


Figure 9. Switching signals generation

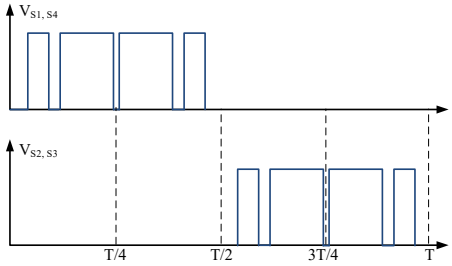


Figure 10. Sample switching signals for N=4

5. Simulation Results

To present the effectiveness of FL based SHE, a model of the system is implemented for the simulations using Matlab/Simulink.

The Simulink block is given in Figure 11 where the aim is to eliminate 10 voltage harmonics (3, 5, 7, 9, 11, 13, 15, 17, 19, 21). The switching angles and the corresponding switching times for N=11 and m=0.9 are given in Table 2.

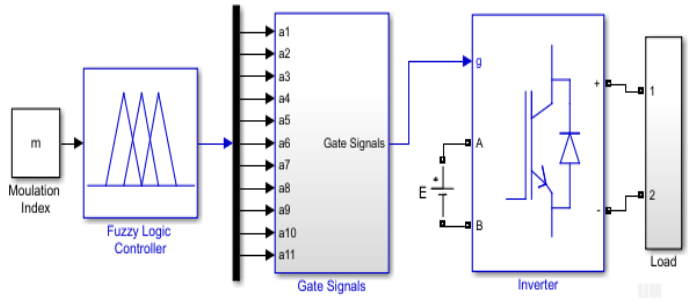


Figure 11. Simulink block of FL based SHE

Table 2. Switching angles (deg) and switching instants (ms) for n=11 and m=0.9

α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}
12.62	15.71	25.38	31.44	38.41	47.25	51.91	63.25	66.15	79.78	81.66
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}
0.70	0.87	1.41	1.74	2.13	2.62	2.88	3.51	3.67	4.43	4.53

For better visualization, a period of the waveform is shown in Figure 12 where horizontal axis is converted to degrees.

Figure 13 shows the inverter output voltage and filtered current waveform when N=11 and m=0.9.

The spectrum of the voltage is presented in Figure 14. The harmonics 3, 5, 7, 9, 11, 13, 15, 17, 19, 21 are all eliminated from this spectrum.

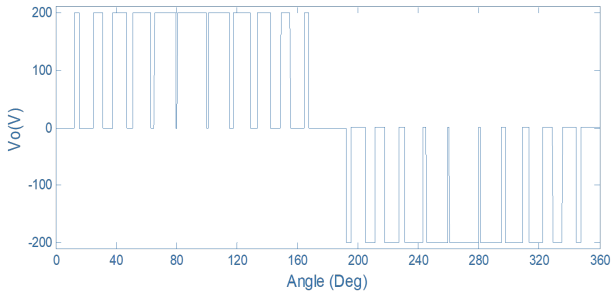


Figure 12. One period of inverter output voltage

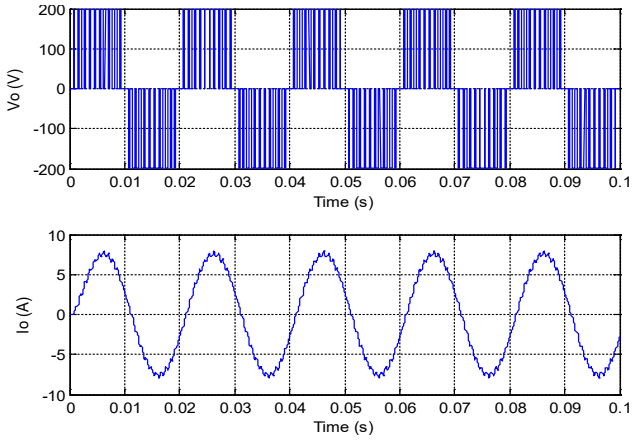


Figure 13. Inverter output voltage and current

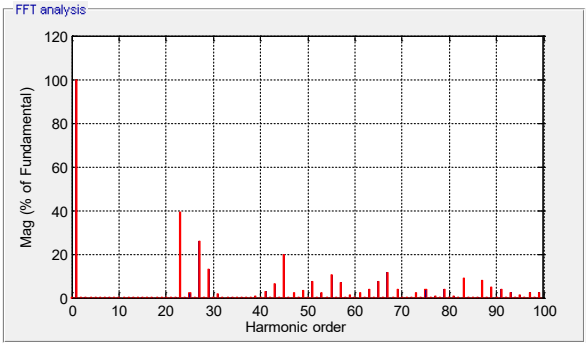


Figure 14. Inverter output voltage and its spectrum

A change in m from 0.8 to 1 is applied at time t=0.1s and the voltage waveform obtained is presented in Figure 15. The output voltage spectra are given in Figure 16 and Figure 17 for the cases m=0.8 and m=1 respectively. Figure 18 is given to clearly see the effect of this change in the voltage waveform. 10 low rank harmonics are eliminated from the spectra.

To show the effectiveness of the presented FL based harmonic elimination system, the first fifteen harmonics which are 3, 5, 7, 9, 11, 13, 15, 17, 19, 21,23, 25, 27, 29, and 31st harmonics are taken to be eliminated. The output voltage waveform and its spectrum are given in Figure 19 and Figure 20. It is clearly seen that 15 predefined harmonics are eliminated.

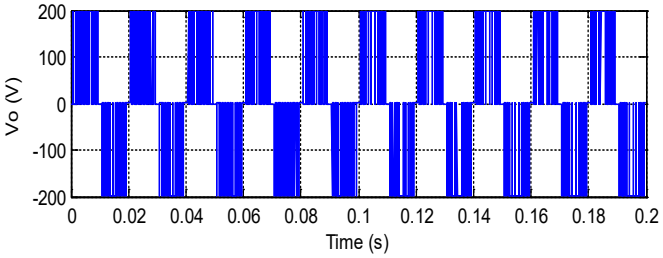


Figure 15. Zoomed output voltage waveform when m changes from 0.8 to 1 at t=0.1

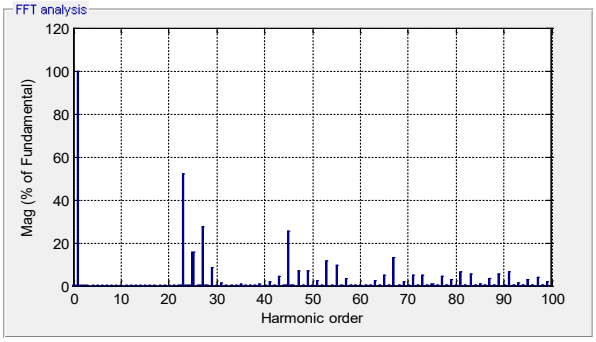


Figure 16. Inverter output voltage and its spectrum for m=0.8 (t<0.1)

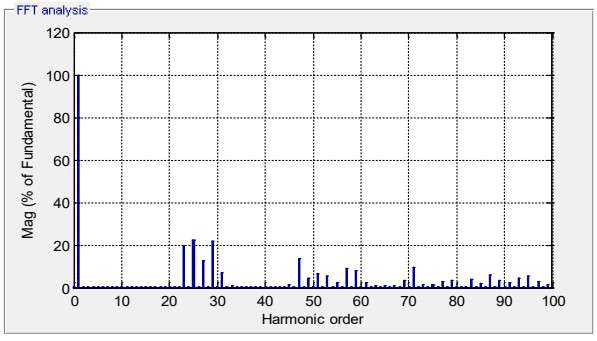


Figure 17. Inverter output voltage and its spectrum for m=1 (t>0.1)

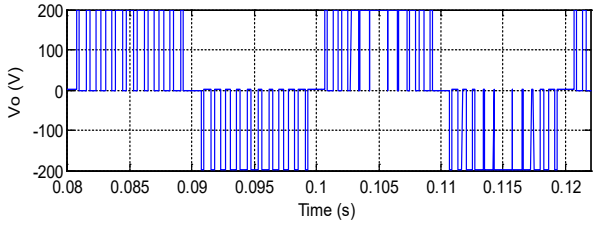


Figure 18. Zoomed inverter output voltage

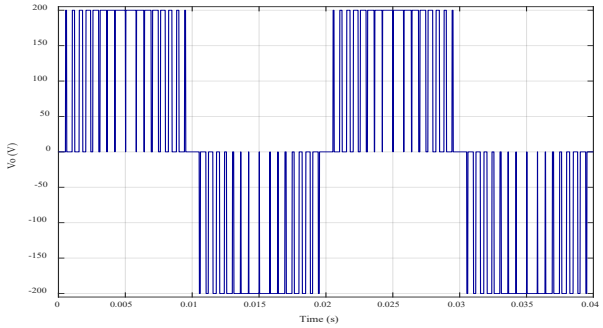


Figure 19. Inverter output voltage waveform with elimination of 15 harmonics

The total harmonic distortion was also examined with the FL based SHE method. THD values are 18.52% and 13.71% for the elimination of 10 harmonics and 15 harmonics. The same load is used in both situations. By using this technique THD is also reduced.

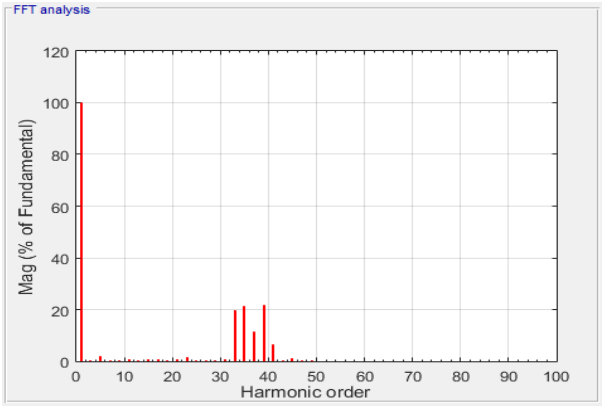


Figure 20. Inverter output voltage spectrum with elimination of 15 harmonics

6. Conclusion

The output voltage of inverters is nonsinusoidal because of the nonlinear switching characteristics of the switches. This voltage contains harmonics with high magnitude with frequencies near fundamental. Distorted voltage used in electronic equipment may reduce the power quality and cause defects. Thus the output voltage waveform of inverters need to be improved. In this study, a FL based SHE for single phase inverters is presented. FL system is used for the computation of switching angles using modulation index. The switching signals are generated for the switches of the inverter using the output of the fuzzy system.

Matlab/Simuink programming environment is used to implement the presented SHE for single phase pwm inverter. Various modulation indices are used for the simulations. Some results are given to show the effectiveness of the technique. The simulation results verified the elimination of the selected harmonics from the frequency spectrum of the voltage waveform. Total harmonic distortion is also reduced by using the presented technique.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] Z. B. Duranay, H. Guldemir, "Fuzzy Logic Based Harmonic Elimination in Single Phase Inverters" in XXVI International Scientific Conference Electronics, Sozopol Bulgaria, 2017. https://doi.org/10.1109/ET.2017.8124342

[2] O. Bouhali, M. Berkouk, B. Francois, C. Saudemont, S. Labiod, "Solving harmonics elimination problem in three-phase voltage controlled inverter using artificial neural networks" Journal of Electrical Systems, 1(1), 39-51, 2005.

[3] M. H. Rashid, Power Electronics Handbook: Devices, Circuits and Applications, Academic Press, 2010.

[4] N. Mohan, T. M. Undeland, W. P. Robbins, Power Electronics: Converters, Applications and Design, John Wiley & Sons Inc., 1995.

[5] E. Hendawi, "Single phase inverter with selective harmonics elimination pwm based on secant method" Int. Journal of Engineering Inventions, 4(12), 38-44, 2015.

[6] M. I. Jahmeerbacus, M. Sunassee, "Evaluation of Selective Harmonic Elimination and Sinusoidal PWM for Single-Phase Dc to Ac Inverters Under Dead-Time Distortion" in 23rd International Symposium on Industrial Electronics, Istanbul Turkey , 2014. https://doi.org/10.1109/ISIE.2014.6864658

[7] B. M. Saied, Q. M. Alias, A. S. AL-soufy, "Intelligent systems based selective harmonic elimination (she) for single phase voltage sourse inverter" Al Rafdain Enginering Journal, 16(3), 71-81, 2008.

[8] P. N. Enjeti, P. D. Ziogas, J. F. Lindsay, "Programmed pwm techniques to eliminate harmonics: a critical evaluation" IEEE Trans. Ind. Appl., 26(2), 302-316, 1990. https://doi.org/10.1109/28.54257

[9] M. S. A. Dahidah, V. G. Agelidis, "Non-symmetrical Selective Harmonic Elimination PWM Techniques: The Unipolar Waveform" in 38th IEEE Power Electronics Specialists Conference, Orlando FL USA, 2007. https://doi.org/10.1109/PESC.2007.4342290

[10] H. S. Patel, R. G. Hoft, "Generalized techniques of harmonic elimination and voltage control in thyristor inverters: part I harmonic elimination" IEEE Trans. Ind. Appl., IA-9(3), 310-317, 1973. https://doi.org/10.1109/TIA.1973.349908

[11] S. Tuncer, Y. Tatar, H. Guldemir, "A shepwm technique with constant v/f for multilevel inverters" Journal of Polytechnic, 8(2), 123-130, 2005.

[12] J. N. Chiasson, M. Tolbert, K. J. McKenzie, Z. Du, "A complete solution to the harmonic elimination problem" IEEE Trans. Power Electron., 19(2), 491-499, 2004. https://doi.org/10.1109/TPEL.2003.823207

[13] V. G. Agelidis, A. I. Balouktsis, M. S. A. Dahidah, "A five-level symmetrically defined selective harmonic elimination pwm strategy: analysis and experimental validation" IEEE Trans. Power Electron., 23(1), 19–26, 2008. https://doi.org/10.1109/TPEL.2007.911770

[14] F. G. Turnbull, "Selected harmonic reduction in static dc-ac inverters" IEEE Trans. Comm. Electron., CE-83(73), 374-378, 1964. https://doi.org/10.1109/TCOME.1964.6541241

[15] D. Czarkowski, D. V. Chudnovsky, G. V. Chudnovsky, I. W. Selesnick, "Solving the optimal pwm problem for single-phase inverters" IEEE Trans. Circuits-I, 49(4) , 465-475, 2002. https://doi.org/10.1109/81.995661

[16] J. Sun, I. Grotstollen, "Pulse Width Modulation Based on Real-Time Solution of Algebraic Harmonic Elimination Equations, Industrial Electronics" in 20th International Conference on Control and Instrumentation, Bologna Italy, 1994. https://doi.org/10.1109/IECON.1994.397753

[17] G. Carrara, S. Gardella, M. Marchesoni, R. Salutari, G. Sciutto, "A new multilevel pwm method: a theoretical analysis" IEEE Trans. Power Electron., 7(3), 497–505, 1992. https://doi.org/10.1109/63.145137

[18] D. Ahmadi, K. Zou, C. Li, Y. Huang, J. Wang, "A universal selective harmonic elimination method for high-power inverters" IEEE Trans. Power Electron., 26(10), 2743–2752, 2011. https://doi.org/10.1109/TPEL.2011.2116042

[19] L. A. Zadeh, "Fuzzy sets" Information and Control, 8, 338-353, 1965. https://doi.org/10.1016/S0019-9958(65)90241-X

[20] A. Kaur, A. Kaur, "Comparison of Mamdani-Type and Sugeno-Type fuzzy inference systems for air conditioning system" Int. J. of Soft Comp. and Eng., 2(2), 323-325, 2012.

[21] H. Singh, M. M. Gupta, T. Meitzler, et al. "Real-life applications of fuzzy logic" Advances in Fuzzy Systems, 2013, 1-3, 2013. http://dx.doi.org/10.1155/2013/581879.

[22] P. M. Larsen. "Industrial application of fuzzy logic control" International Journal of Man-Machine Studies. 12(1), 3-10, 1980. https://doi.org/10.1016/S0020-7373(80)80050-2.

[23] R. R.Yager, and L. A. Zadeh. "An introduction to fuzzy logic applications in intelligent systems". Vol. 165. Springer Science & Business Media, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 161-165 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Implementation of a Robot Contest for Distance Education

Eiken Lübbers*, Dierk Schoen

Wilhelm Buechner University for Applied Sciences, 64319, Germany

ARTICLEINFO

Article history:
Received: 10 August, 2018
Accepted: 24 August, 2018
Online: 29 September, 2018

Keywords:
Robot Sumo
Sensorics
Distance Education

ABSTRACT

Robot contests are suitable for bachelor courses to gather hands-on experience in engineering diciplines such as mechanics, electronics and embedded software. After summarizing a concept of a robot contest for distance education (robot sumo) as presented in earlier publications this paper describes the implementation and first results of this concept. The contest is seperated into different tasks such as line detection and opponent location. As an example for transferring theoretical knowledge into a working prototype a simple sensor design for robot sumo is described and tested. Additionally a first student concept of a self designed robot is presented.

1 Introduction

This paper adds first implementation experiences to a robot contest concept originally presented at the 2017 International Conference on Research and Education in Mechatronics (REM) [1]. There it is shown that contests play an important role in education and in introduction of higher technology into the bachelor course [2], [3]. The presented concept is suitable for use in distance education where on-site laboratories are not readily available. Out of the evaluation of different contest formats robot sumo is selected as best suited. As an addition to this concept, that will be briefly summarized, this paper will describe the implementation of the contest into the curriculum at the Wilhelm Büchner University of Applied Sciences (WBU). To participate in a contest of robot sumo it is necessary to implement different strategies into a robot. To help students create their own solutions WBU has selected a platform compatible to the Arduino standard. Together with some prototyping parts this platform is given to the students during the project phase in the 7th semester in the module "Project Study for Engineers". The different solution out of these projects will enter into a future contest at the WBH where they will compete against standard implementations that can be purchased. To illustrate the use of creative solutions an innovative sensor implementation is shown to locate and range the opponent. Measurement data and tests are shown to evaluate the performance of the sensors. Finally, a robot is presented that was developed and made by a student during his bachelor thesis. This robot is entirely made by 3D-printing and shows, what can be achievied within the three months avalible for both the bachelor thesis and the project study.

2 Contest Concept

The original paper [1] evaluated different robot contest formats. The criteria were robot size, arena size, budget, complexity, time to completion and research potentials. Robot Sumo in the mini class has been selected as it seems the most suitable for table top and home development [1].

3 Robot Sumo

The contest is based on the traditional Japanese game of sumo wrestling (jap. 相撲 Sumō) where two contestants try to push each other out of a circular ring (jap. 土俵 Dohyō). The contestants start behind two lines in the middle of the ring (jap. 仕切り線 Shikirisen = Shikiri lines) [4].

There are 28 articles in the unified sumo rules making this a well designed contest [5]. Figure 1 is showing a typical Sumo match situation where one robot is pushed over the edge and touches the ground outside the ring and thus loses this round.

In order to compete in a robot sumo contest the robot must implement different strategies. These are often divided into the different domains of mechanical, electrical/electronical, and software. There are three major strategies to a robot sumo match:

*Wilhelm Buechner University for Applied Sciences, 64319, Germany, Email: eiken.luebbers@wb-fernstudium.de

https://dx.doi.org/10.25046/aj030520

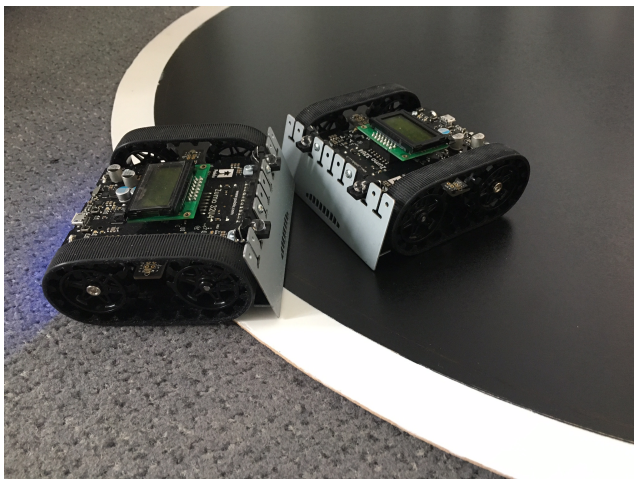


Figure 1: Typical Sumo match situation
(picture: Eiken Luebbers)

3.1 Do not leave the ring

The robot that leaves the ring first will loose the round. Thus it is important to detect the ring border so that the robot will not leave the ring by itself while moving. The rules ask for a black surface of the ring and a 2.54 cm white line around the border (mini class). The most common strategy is to detect the change in reflectance between the black and white surfaces and move backward or turn to avoid leaving the ring. Reflectance sensors are inexpensive and most will provide a digital signal that can be easily processed by microcontrollers.

3.2 Locate your opponent

The most important strategy is to locate the opposing robot so that it can be targeted with the actors. The solutions vary from moving randomly around the ring until an acceleration sensor registers a collision to LiDAR-Sensors that will determine the exact location and range. All strategies aim to put the opponent in front of the own robot so it can be pushed out of the ring. A simple but efficient implementation of this strategy will be described later in section 6.

3.3 Push the opponent out of the ring

This strategy seems to be the most straightforward implementation: Move in the direction of the located opponent. But it can be subdivided into mechanical parts such as blades, that can lift the opponent so it will loose traction and motors that could combine traction and momentum. The mechanical and electromechanical design will be part of the outlook, a first implementation can be seen in section 7.

4 Robot platform

Not all students have the experience, time, or knowledge to design a complete robot system. To concentrate on the chosen strategy in the project a platform is provided by the WBH to help the students with an easy start into development. First experiences show that there are some groups that concentrate on software, some on sensors and others on mechaical design.

The platform is a commercially available robot specifically designed for the robot sumo mini class. It is an electromechanical platform that includes the power supply out of four AA-batteries, downward-oriented reflectance sensors for the ring edge detection, general purpose inertial sensors, the motor drivers including metal-gear motors, and a differential rubber chain drive. It also includes an Arduino-compatible interface to add a microcontroller and/or sensors.

When students sign up for the project study (see section 5), the WBH will provide two of these platforms together with two Arduino compatible microcontroller-boards, two mini-breadboards for rapid prototyping and a number of wires with headers for quick installation of sensors. No soldering is necessary to get started with the platform and first implementations.

5 Project Study for Engineers

The project study has been identified as the proper format for the robot sumo development in [1]. The students learn project management principles and team work. And it must have a measurable outcome within a 3 month time frame. When working on a robot during the module "Project Study for Engineers" the students gain practical knowledge in mechatronic systems. The robots need mechanics, electronics and software to function. Students need to research the current technologies and methods and implement them into the contest. Furthermore, the project needs to be managed in time and within a given budget. This adds economic sciences and industrial engineering to the contest. And finally the project needs to be productive for a live contest at a certain interval, adding working under pressure to the picture. All this teaches the students to be productive in future jobs.

6 Sensor Solution

To demonstrate a creative solution for the strategy of "locate your opponent" the implementation in another commercially available platform is analyzed. This platform adds a microcontroller and infrared proximity sensors to the platform described in section 4. It combines an inexpensive standardized electronic hardware with a non-standard control by software to combine detecting and ranging of an object.

6.1 Sensor Electronics

The basis for the hardware is the digital proximity sensor TSSP77038 (Vishay Semiconductors) as it is used e.g. in automatic water fountains [6]. The sensor has a digital output active-low interface for microcontrollers

to signal if an object is detected or not. The sensor is sensitive for infrared light of a certain bandwidth, the typical wavelength is 940 nm. The sensor eliminates interferences from natural or artificial light by detecting only modulated infrared light with a frequency of 38 kHz. To illuminate an object that can be detected, an infrared light emitting diode (IR-LED) is driven with a carrier frequency of the same 38 kHz. If enough modulated infrared light is reflected from an object onto the sensor, the output is driven to low to signal an object in range of the sensor.

To add a relative range detection the light amount of the emitting diode is varied using a pulse-width-modulation (PWM) as an envelope function. This allows to control the output power of the diode and thus the amount of reflected light. If an object is detected with the full power of the diode and the power is then stepwise reduced, a relative range of the detected object can be determined. The amount of power when the object is no longer detected is a measure for distance. The less power is needed, the closer the object is. If used with a single emitter and single sensor a one dimensional measurement is possible.

The typical information needed for the strategy "push the opponent out of the ring" is to know, if the robot needs to turn left or right to face the opponent directly with its front so when driving forward, it will hit the opponent an can push it in front of it. To be able to have this twodimensional information the robot needs at least two emitter/sensors combinations.

The implementation actually uses three sets of emitter and sensor. One facing front, one facing left, and one facing right. Each set is controlled individually by the software.

6.2 Control Software

Each infrared light emitting diode can be individually controlled by software via PWM to send out a certain amount of light modulated with 38 kHz. Each sensor gives back the information, if enough modulated light in its field of view has been detected. Sending only light to the left will illuminate objects that are left of the robot. The right sensor should "see" significantly less light than the left or front sensor, depending on the position of the reflecting object.

6.3 Measurement Results

To test the strategy a white object was placed at certain angles and distances to the robot. The relative distances of each sensor where recorded to se if this concept can detect direction and distance of an object.

The PWM can be controlled in 6 levels, where 1 has the highest power output and 6 the lowest. The object was placed in 20 degree intervals around the center of the robot. The results are shown in figures 2, 3, and 4.

Both upper graphs show that the range can be clearly distinguished. Also it can be seen that each sensor has a certain field of view. The bottom graph shows the overall detection capability of a white object placed at a 10.5 cm distance. But it also shows that there are certain blind spots between 40 and 60 degrees each left and right of the front. The blind spot at the rear is expected as there are no LEDs or sensors here.

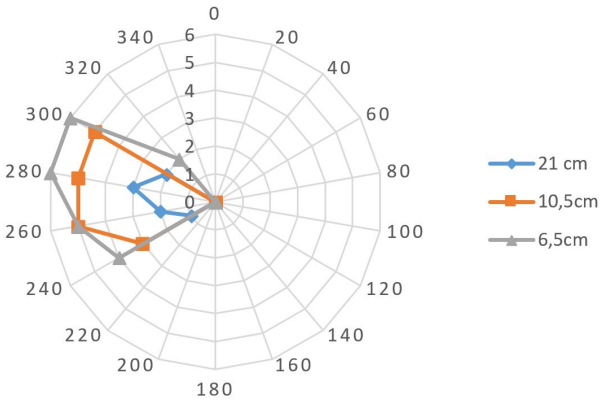


Figure 2: Measurements results left LEDs/left sensor

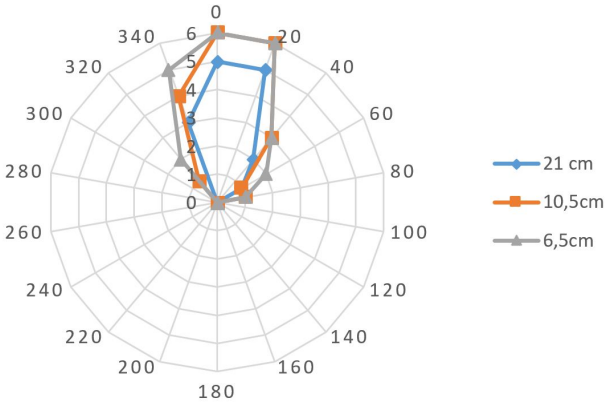


Figure 3: Measurements results
middle: right LEDs/front sensor

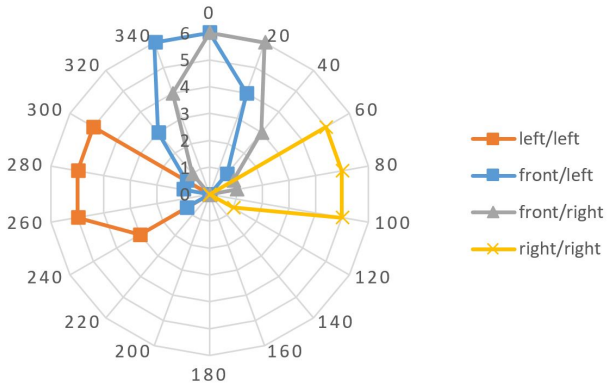


Figure 4: Measurements results
bottom: all LEDs/sensors

It was shown, that the sensor concept can detect relative distance and direction of the opponent. When the front sensor shows the same distance when left and right LEDs are used, it can be assumed, that the obejct or opponent is directly in front of the robot (see figure 4 bottom graph at 0 degrees). This information can then be used in the next step where the motors are controlled to move the robot in the right direction. If the detection logic is used continuously the robot can also follow moving opponents.

7 Student Robot Solution

After showing a possible solution with an existing hardware it requires further consideration if it is possible for students to actually complete a robot construction within 3 months that are available for the project study. A student at the WBH has therefore been given the task to design and build a complete robot system with the use of a 3D-printer.

The first task was to analyze potential techniques for sumo robots [7]. The fields of traction (e.g. 2 wheel drives vs. 4 wheel drives), shields and shovels (for protection or to lift the opponent), ropes (to keep other robots at distance), flags (to trick the opponent to follow the flag) or size (keep below the line of sight of the opponent) were analyzed. Additionally the sensor technology was researched. Inertial sensors, several distance sensors, reflection sensors and cameras were described. The research was completed with a look into the Arduino platform and 3D-printing.

For the actual prototype an architecture was developed and the proper components were selected. For the traction a rubber chain differential drive was chosen. This allows the robot to distribute all its weight onto driven wheels or chains respectively to have the maximum traction within the allowed weight range. The force with which the robot can push ("traction") is the frictional force that the driven wheels can generate. This is proportional to the weight of the robot (normal force) at the contact between driven wheel and ring surface (Amontons' First Law). If the robot had only two driven wheels and one or two support wheels then some of the weight rests on the passive wheel(s) and cannot support the friction of the active drives. A chain drive is easier to implement as 4 wheel drive and still distributes all weight onto the driven components.

Out of his research the student chose Time-of-Flight (ToF) laser sensors. These sensors determine the distance to an object by measuring the time a laser pulse needs to travel to an object and back. The sensor used is the "world's smallest ToF sensor" from STMicroelectronics [8]. It is a sensor system in one housing combining a laser diode with a detection circuit using an single photon avalanche diode array and an on-silicon processing unit. It communicates via serial interface with a host controller.

The mechanical parts are 3D-printed. Figure 5 shows the finished robot. Motors and gears were purchased as were the electronic components. In total the price of the robot was about half compared to the commercial product while having more flexibility in design, sensors, and computing power. The shown robot uses two arduino controllers boards that process sensors and mechanics seperately and communicate with each other via an onboard network.

8 Summary and outlook

In addition to the original paper presented at the 2017 International Conference on Research and Education in Mechatronics (REM) [1] where the concept of robot sumo for distance education is explored this paper adds some practical measurements and a student solution to the concept. The first competition will add a major milestone to this concept of learning. Students will have completed a project where the robot is able to compete in robot sumo. Different strategies in terms of mechanics and tactics will be the source of future publications. After showing what is possible for one student within 3 months the project was oficially offered at WBH for the project study in the bachelor courses in Spring of 2018. The project has been promoted by newsletter and YouTube-videos [9]. Since launch the commercial platform has been supplied to a number of teams that are now designing and testing their robots. The first competition will be held at the end of 2018 when at least 4 teams have completed the project. The competition will be supplemented by 4 standard robots that are comercially available. The competition will be recorded and published on youtube.

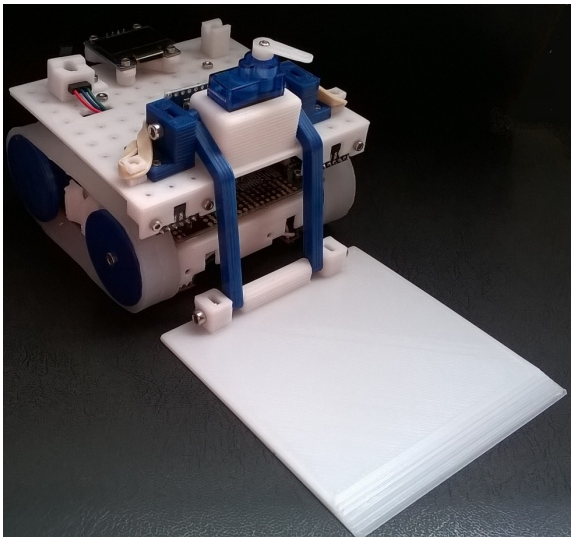


Figure 5: CAD parts and complete 3D-printed robot (source: [7])

References

[1] E. Lübbers, D. Schoen, "Robot contest concept for distance education" in 18th International Conference on Research and Education in Mechatronics (IEEE), Wolfenbüttel Germany, 2017. https://doi.org/10.1109/REM.2017.8075253

[2] T. Verhoeff, "The role of competitions in education" in Future World - International Conference & Exhibition, 1997. [Online]. Available: http://olympiads.win.tue.nl/ioi/ioi97/ffutwrld/competit.pdf

[3] A. Kolmos, "Problem-based and project-based learning" in University Science and Mathematics Education in Transition, O. Skovsmose, P. Valero, and O. R. Christensen, Eds. Berlin: Springer-Verlag, 2009, pp. 261–280.

[4] Wikipedia, "Sumō", 10.05.2017. [Online]. Available: https://de.wikipedia.org/w/index.php?oldid=165000523

[5] Robotics Society of America, Inc, "Unified sumo robot rules", 2015. [Online]. Available: http://robogames.net/rules/all-sumo.php

[6] Vishay Semiconductors, "IR Sensor Module for Reflective Sensor, Light Barrier, and Fast Proximity Applications", 2018. [Online]. Available: https://www.vishay.com/docs/82470/tssp770.pdf

[7] A. Kusmin, "Entwicklung einer autonomen Roboter-Plattform für Roboter-Sumo", Bachelor Thesis, Wilhelm Büchner University of Applied Sciences Pfungstadt Germany, 2017.

[8] STMicroelectronics, "World smallest Time-of-Flight (ToF) ranging sensor" 2018. [Online]. Available: https://www.st.com/content/st_com/en/products/imaging-and-photonics-solutions/proximity-sensors/vl53l0x.html

[9] YouTube, "Roboter-Sumo - Wilhelm Büchner Hochschule", 2018. [Online]. Available: https://youtu.be/8qtDo-sM_0s

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 166-170 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Influence of Torrefaction on Gasification of Torrefied Palm Kernel Shell

Razi Ahmad[*,1,4], Mohd Azlan Mohd Ishak[1], Khudzir Ismail[2], Nur Nasulhah Kasim[3]

[1]*Faculty of Applied Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

[2]*Faculty of Applied Sciences, Universiti Teknologi MARA, Campus Arau, 02600 Arau, Perlis, Malaysia*

[3]*Coal and Biomass Energy Research Group, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia*

[4]*School of Environmental Engineering, Universiti Malaysia Perlis, 02600, Arau, Perlis, Malaysia*

ARTICLE INFO

Article history:
Received: 09 August, 2018
Accepted: 14 September, 2018
Online: 27 September, 2018

Keywords:
Palm kernel shell
Torrefaction
Pretreatment
Gasification

ABSTRACT

In this study, torrefaction pretreatment on palm kernel shell (PKS) was investigated using fixed bed reactor. The PKS was torrefied at the temperatures of 210, 230, 250, 270 and 290 °C. The characteristics between untreated and torrefied PKS were compared. The results showed that, the mass and energy yield lessened, while the calorific value augmented with the increasing torrefaction temperature. Furthermore, with the rise of temperature, the oxygen composition, O/C ratio, oxygenated compounds and volatile matter of torrefied PKS decreased, but, the carbon and fixed carbon content increased. The composition of carbon in torrefied PKS was toward coal where equivalent calorific values was achieved. The gasification of torrefied PKS enhanced the product yield which produced higher gas, lower tar and char yield than the gasification of untreated PKS. Gasification of torrefied PKS increased the gas yield by 16.9 % than the untreated PKS. The tar and char yield of torrefied PKS decreased by 19.4 % and 25.9 %, respectively than the untreated PKS. Therefore, the torrefied PKS, by which their physical and chemical properties have been improved through torrefaction pretreatment is more suitable to be used in gasification and co-gasification as their influences are significant than the untreated PKS.

1. Introduction

Nowadays, the growing utilization of energy, worries on the worldwide environmental difficulties and lessening of fossil fuel, lead the nation to head for clean and renewable energy. This paper is an addition of the work formerly presented in 4th IET Clean Energy and Technology Conference 2016 [1]. Among of the main problems for fossil fuels are the discharge of contaminants such as carbon dioxide, sulfur and nitrogen oxide towards environment [2]. Therefore, biomass is one of the most attractive and broadly used renewable energy source, become important as an alternative energy resource due to little sulfur composition and neutral CO_2 supply [3].

Despite the great prospective of biomass, it has the drawbacks on its properties such as high moisture content, low energy density and hydrophilic characteristics [4-5]. Thus, these characteristics of biomass fuel are connected with some complications in biomass thermal conversion such as in gasification. Previous studied [6-7], revealed that high oxygen compound in biomass lower the gasification productivities compare with less oxygen, for example coal. Thus, alteration the properties of biomass preceding gasification is necessary.

A pretreatment step preceding to thermal conversion is required in the direction to reduce some of the aforementioned problems. Thus, torrefaction appears to be an effective route. Torrefaction involves pretreatment at temperature ranges of 200 to 300 °C in atmospheric surrounding. The pretreated biomass formed a fuel with low moisture and great energy content [2]. Previous studies also show other advantages of this torrefaction pretreatment, such as improving feedstock hydrophobicity, homogeneity and grindability [8-9].

Palm as the highest provider to biomass incomes in Malaysia has appealed huge consideration to achieve the renewable energy demands [10]. In 2016, Malaysia produced 4.19 MnT of PKS, as

[*]Razi Ahmad, Faculty of Applied Sciences, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, +6049798626, razi@unimap.edu.my

residues from oil palm industry [11]. Thus, transforming PKS to bio-fuel under a thermal conversion offers a greater advantage to substitute fossil fuels, and it minimizes the disposal problems related with the generation of agricultural by-products [12]. PKS exhibited great prospective as fuel to produce gas with enhanced hydrogen and energy content [13]. However, high moisture, low heating value and energy density inhibit the PKS as valuable fuel [14]. Accordingly, these complications can be handled through torrefaction.

Consequently, the research objective was to explore the influence of torrefaction temperature on the characteristic of torrefied PKS. Further, the gasification of torrefied PKS was investigated.

2. Methodology

2.1. Materials

PKS sample was obtained from oil palm factory in Penang, Malaysia. It was crushed and sieved to get particle sizes between 200 to 400 µm. The inherent moisture was removed by drying the sample in an oven for 24 hours at 105 °C.

2.2. Pretreatment

Torrefaction, which is a mild pyrolysis pretreatment was studied thru a fixed-bed reactor at an atmospheric pressure. The reactor has inner diameter and height of 0.06 m and 0.3 m, respectively. The electric furnace surrounding reactor tube was used to heat the reactor. Figure 1 displays a schematic diagram of the pretreatment system.

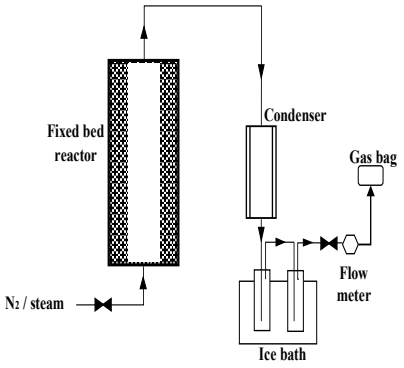


Figure 1. Schematic diagram of the reactor set-up

The sample weight of 5 g was positioned inside the reactor. The reactor was flowed with 500 ml/min nitrogen gas to generate an inert atmosphere. The reactor temperature was fixed at reaction temperatures (210, 230, 250, 270 and 290 °C) with heating rate and retention time of 10 °C/min and 1 hr, correspondingly. The torrefied PKS was weighted when the process was completed and reached the room temperature.

Mass and energy yield are main factors in the pretreatment. Energy yield represented the quantity of energy preserved after the pretreatment and is measured using mass yield of final product. Torrefied biomass has fewer energy yield than untreated biomass, due to reduce in volatile matter of torrefied biomass [15]. The mass yield (Y_m) and energy yield (Y_e) were examined via equations (1) and (2), correspondingly.

$$Y_m = (M_t / M_u) \times 100 \quad (1)$$

$$Y_e = Y_m \times (HHV_t / HHV_u) \quad (2)$$

where M_u = mass of untreated biomass, M_t = mass of torrefied biomass, HHV_u = heating value of untreated biomass and HHV_t = heating value of torrefied biomass.

2.3. Characterization

The elemental composition (C, H, N, S and O) was examined using CHNS-O elemental analyser. The proximate analysis was inspected via Mettler Toledo thermogravimetric analyser. The calorific value (CV) was measured using Leco bomb calorimeter. The functional groups were discovered using Perkin Elmer fourier transform infra-red (FTIR) spectroscopy. Table 1 listed the properties of untreated PKS.

Table 1. Properties of untreated PKS

Analysis	Value
Elemental composition (wt. %)	
Carbon	47.7
Hydrogen	5.5
Nitrogen	0.4
Sulfur	0
Oxygen[a]	46.4
Proximate analysis (wt. %)	
Moisture	10.6
Volatile matter	77.5
Ash	0.9
Fixed carbon	11.0
Calorific value (MJ/kg)	18.2

[a] By different

2.4. Gasification Experiment

Figure 1 displays the gasification system of PKS. The sample weight of 5 g was positioned inside the reactor. A nitrogen gas was flowed to the reactor for 10 min formerly the test. The sample was gasified at gasification temperature (800 °C) with heating rate of 50 °C/min. The nitrogen flow of 500 ml/min was continued to generate an inert condition. After the temperature of 800 °C had reached, the steam was streamed into the reactor and the nitrogen flow was stopped. The steam gasification of the sample was held for 45 min.

The volatile product and steam which left the reactor from the upper side were condensed in a tar trap. The solid residue was weighted as char. Tar yield in the tar trap was measured. The gas product was inspected using changed of total mass balances. The gasification was repeated for verification of the outcomes.

3. Result and Discussion

3.1. Mass and Energy Yield

Figure 2 presents the mass and energy yield of torrefied PKS under different torrefaction temperatures. The mass and energy yield reduces by increasing the temperature. The mass yield ranges from 88 to 65 % of torrefied PKS at temperature ranges of 210 to 290 °C. This displays that the conversion of PKS was increased from 12 to 35 %. The slight conversion at the temperature of 210 °C was reflected to the loss of moisture. Thus, the PKS torrefaction was insignificant at low temperature. At upper temperature between 230 to 290 °C, mass reduction was due to the major

hemicelluloses and minor lignin decomposition [4]. Some authors [16-17] established that the main decomposition part during torrefaction was hemicelluloses.

The energy yield of torrefied PKS was considerably reduced to 75 % at pretreatment temperature of 290 °C. This observation was mostly due to the additional decomposition of cellulose. Accordingly, more than 75 % of energy yield was able to be reserved at the pretreatment temperature between 250 to 270 °C. Hence, torrefaction of PKS above 290 °C is not suggested in order to avoid the loss of energy yield below 75 %.

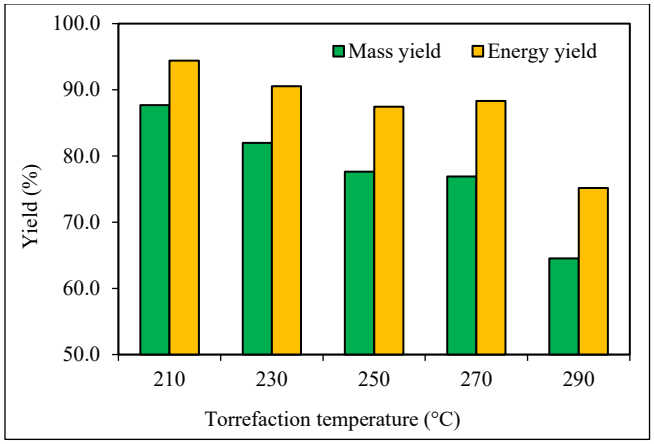


Figure 2. The influence of torrefaction temperature on mass and energy yield

3.2. Fixed Carbon, Volatile Matter and Calorific Value

Figure 3 presents the influence of pretreatment temperature on fixed carbon content and volatile matter of torrefied PKS. The fixed carbon content of torrefied PKS increased while volatile matter decreased notably with the rise in torrefaction temperature. At high temperature (290 °C), the fixed carbon of the torrefied sample improved above 50 % with comparison to the untreated sample. The torrefied sample showed huge reduction of volatile matter with close to 50 % with increasing reaction temperature up to 290 °C. The hemicellulose content in PKS is easy to degrade during torrefaction process. The results on the extensive volatile matter reduction were comparable to the work published by Uemura et al. [18], Matali et al. [19] and Sabil et al. [20] in their study of agricultural wastes.

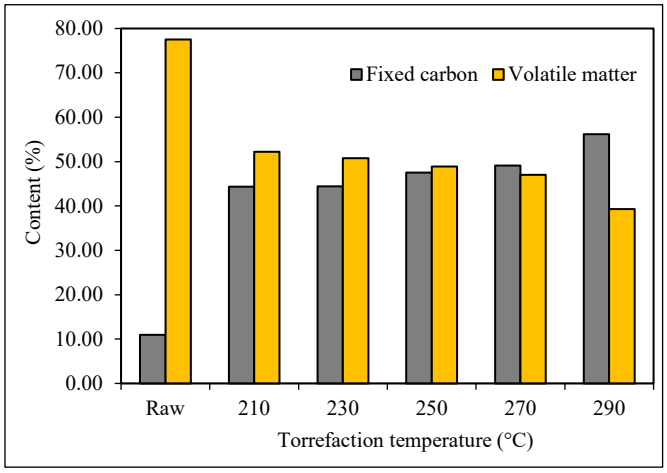


Figure 3. The influence of torrefaction temperature on fixed carbon and volatile matter

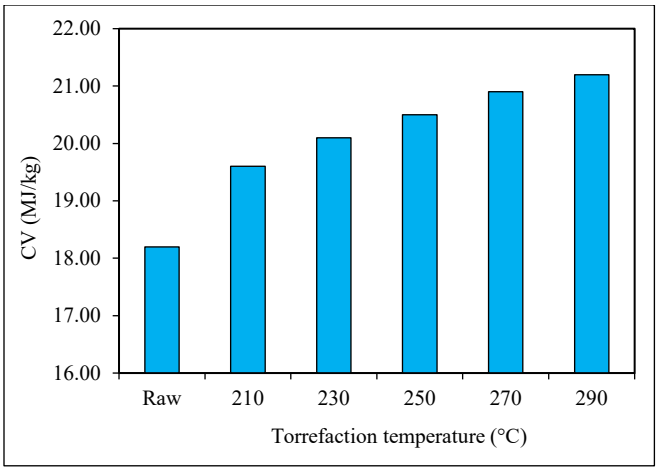


Figure 4. The influence of torrefaction temperature on calorific value

The CV of torrefied PKS is presented in Figure 4. The CV of torrefied PKS increased by rising the temperature. Improvement of CV is related with the rise of fixed carbon component. Accordingly, the PKS energy value enriched with pretreatment.

3.3. Carbon and Oxygen Content

The carbon and oxygen content of torrefied PKS are presented in Figure 5. Overall, the torrefied PKS displayed lower oxygen and higher carbon composition than untreated PKS by increasing the torrefaction temperature. The oxygen was reduced to 39 % and the carbon was increased up to 56 % at the highest pretreatment temperature of 290 °C. These outcomes appear to be in agreement with the earlier reports [21-22]. Moreover, O/C ratio of torrefied PKS reduced by rising the temperature. The reduction of the O/C ratio also indicates the measure of conversion efficiency and oxidation degree of torrefied product [4].

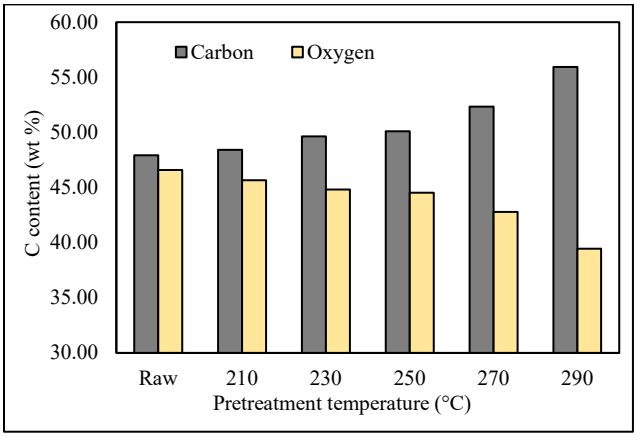


Figure 5. Carbon and oxygen content of torrefied PKS

3.4. Functional Group

Figure 6 shows the chemical structure alteration of PKS samples via FTIR. The spectrum shape was comparable for untreated and torrefied PKS, but the peak strength was dissimilar.

A broad band of 3300 cm^{-1} connected to -OH stretching which related to the alcohols and phenols. The -OH peak reduced

considerably as the pretreatment temperature increased. The aliphatic methylene group was denoted at peak of 2920 cm^{-1}. The C=O bond which is associated with aldehydes, acids and ketones was detected at 1730 cm^{-1}. At greater pretreatment temperature, the peak intensity reduced due to the breakdown of hemicellulose. The C-O stretching and O-H alteration of organic components are assigned in the peak ranges of 1000 - 1500 cm^{-1}. Granados et al. [23] also found the similar trend which the intensity of these peaks were reduced with increasing torrefaction temperature. Peak of 700 cm^{-1} indicated the aromatic groups.

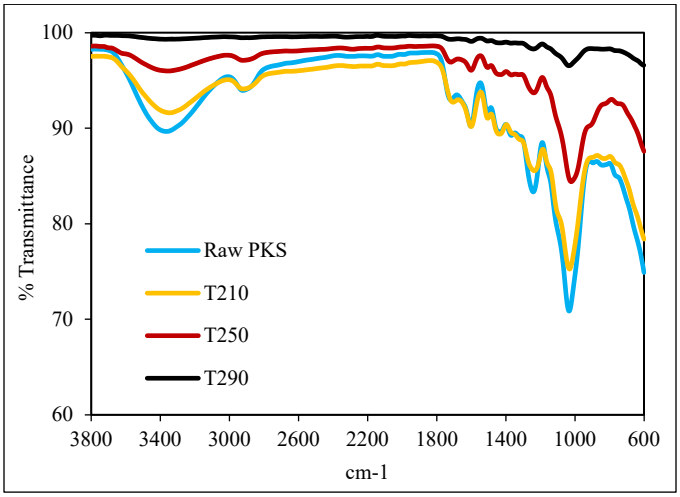


Figure 6. FTIR spectrum of PKS samples

3.5. Effect of Product Yield on Gasification of Torrefied PKS

Figure 7 shows the gasification product yield of untreated and torrefied PKS at gasification temperature of 800 °C. The torrefied PKS produced 16.9 % higher gas yield than the untreated PKS. The torrefied PKS revealed notable influence on the gas production to produce high gas yield compared to untreated PKS. This result was also in agreement with Berrueco et al. [24] which produced higher gas yield using torrefied sample.

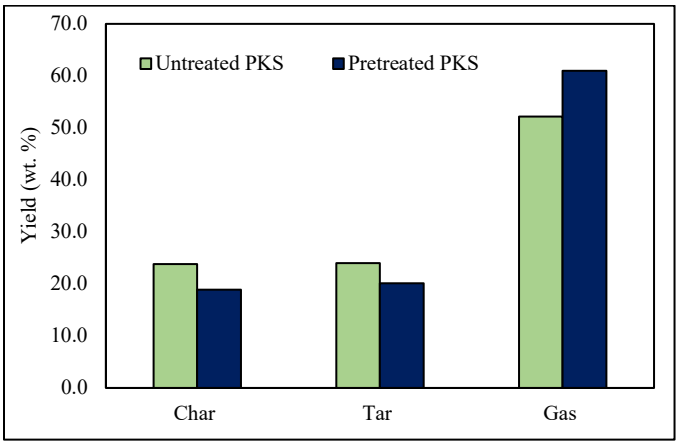


Figure 7. Gasification product yield of untreated and pretreated PKS

The tar yield reduced from 24.0 % to 20.1 % using for the torrefied PKS. The torrefied PKS had markedly reduced the tar yield about 19.4 % as an impact of the partial elimination of oxygenated components and volatiles through the torrefaction. Dudyunki et al. [25] also exhibited a comparable outcome as the effectiveness of using torrefied biomass to produce lower tar than the untreated biomass.

The gasification of torrefied PKS decreased the char yield from 23.8 % to 18.9 %. Low char yield using torrefied PKS was connected with the increased of solid conversion to gas product. Moreover, this event was influenced by the low moisture and oxygenated compound of torrefied PKS.

4. Conclusion

The influences of pretreatment temperature on torrefied PKS was investigated successfully. It was determined that the CV, fixed carbon and carbon content increased, however, mass and energy yield, volatile matter and oxygen content reduced, as the temperature augmented. Furthermore, oxygenated peak intensity in FTIR spectra decreased with increasing temperature. Therefore, PKS revealed a high value biofuel at reaction temperature from 250 °C to 290 °C. PKS torrefied at 250 °C showed significant mass and energy yield around 75 % and 85 %, respectively. The CV also increased more than 10 % compared to untreated PKS. The considerable reduction of oxygenated peak intensity was also found for torrefied PKS at 250 °C.

The gasification with torrefied PKS shows a positive effect in terms of product yield distribution. The torrefied PKS produced 16.9 % higher gas yield than the untreated PKS. The tar yield was reduced from 24.0 % to 20.1 % using the torrefied PKS. Also, the gasification of torrefied PKS decreased the char yield from 23.8 % to 18.9 %. Consequently, the torrefaction pretreatment, which improved the PKS properties, enhanced the gasification performance by producing high gas yield with low tar and char yield.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by Ministry of Education, Malaysia under FRGS Grant (1/2017/TK10/UITM/2/11).

References

[1] R. Ahmad, K. Ismail, M. A. M. Ishak, N. N. Kasim, and C. Z. A. Abidin, "Pretreatment of palm kernel shell by torrefaction for co-gasification," in 4th IET Clean Energy and Technology Conference (CEAT 2016), 2016.. https://doi.org/10.1049/cp.2016.1327 IET Digital Library

[2] C. Berrueco, J. Recari, B. M. Güell, and G. del Alamo, "Pressurized gasification of torrefied woody biomass in a lab scale fluidized bed," *Energy*, 70, 68–78, 2014. https://doi.org/ 10.1016/j.energy.2014.03.087

[3] J. S. Brar, K. Singh, J. Wang, and S. Kumar, "Cogasification of Coal and Biomass: A Review," *Int. J. For. Res.*, 2012, 1–10, 2012. https://doi.org/ 10.1155/2012/363058

[4] W. Chen, J. Peng, and X. T. Bi, "A state-of-the-art review of biomass torrefaction , densi fication and applications," *Renew. Sustain. Energy Rev.*, 44, 847–866, 2015. https://doi.org/10.1016/j.rser.2014.12.039

[5] R. Ahmad, M. Azlan, M. Ishak, N. N. Kasim, and K. Ismail, "Effect of Different Pretreatments on Palm Kernel Shell And Low-rank Coal during Co-gasification," *Progress in Petrochemical Science*, 1–7, 2018.

[6] Z. Chen, S. Zhang and Z. Chen."An integrated process for hydrogen-rich gas production from cotton stalks: The simultaneous gasification of pyrolysis gases and char in an entrained flow bed reactor," *Bioresource Technology*,

198, 586-592, 2015. https://doi.org/10.1016/j.biortech.2015.09.015
[7] A. C. C. Chang, H.-F. Chang, F.-J. Lin, K.-H. Lin, and C.-H. Chen, "Biomass gasification for hydrogen production," *Int. J. Hydrogen Energy*, 36(21), 14252–14260, 2011. https://doi.org/10.1016/j.ijhydene.2011.05.105
[8] H. Nam and S. Capareda, "Experimental investigation of torrefaction of two agricultural wastes of different composition using RSM (response surface methodology)," *Energy*, 91, 507–516, 2015. https://doi.org/10.1016/j.energy.2015.08.064
[9] D. Nhuchhen, P. Basu, and B. Acharya, "A Comprehensive Review on Biomass Torrefaction," *Int. J. Renew. Energy Biofuels*, 2014, 1–56, 2014. https://doi.org/10.5171/2014.506376
[10] Z. A. Z. Pooya Lahijani, "Gasification of palm empty fruit bunch in a bubbling fluidized bed: A performance and agglomeration study," *Bioresour. Technol.*, 102(2), 2068–2076, 2011. https://doi.org/10.1016/J.BIORTECH.2010.09.101
[11] D. Ahmad Kushairi, "Malaysian Oil Palm Industry Performance 2016 and Prospects for 2017," in Palm Oil Econ. Outlook Seminar, 2017.
[12] R. Ahmad, N. Hamidin, U. F. M. Ali, and C. Z. A. Abidin, "Characterization of Bio-Oil From Palm Kernel Shell Pyrolysis," *J. Mech. Eng. Sci.*, 7(12), 1134–1140, 2014. https://doi.org/10.15282/jmes.7.2014.12.0110
[13] N. A. Samiran, M. N. M. Jaafar, J. H. Ng, S. S. Lam, and C. T. Chong, "Progress in biomass gasification technique - With focus on Malaysian palm biomass for syngas production," *Renew. Sustain. Energy Rev.*, 62, 1047–1062, 2016. https://doi.org/10.1016/j.rser.2016.04.049
[14] M. A. Sukiran, F. Abnisa, W. M. A. Wan Daud, N. Abu Bakar, and S. K. Loh, "A review of torrefaction of oil palm solid wastes for biofuel production," *Energy Convers. Manag.*, 149(10), 101–120, 2017. https://doi.org/10.1016/j.enconman.2017.07.011
[15] E. M. Gucho, K. Shahzad, E. A. Bramer, N. A. Akhtar, and G. Brem, "Experimental study on dry torrefaction of beech wood and miscanthus," *Energies*, 8(5), 3903–3923, 2015. https://doi.org/10.3390/en8053903
[16] M. Asadullah, A. M. Adi, N. Suhada, N. H. Malek, M. I. Saringat, and A. Azdarpour, "Optimization of palm kernel shell torrefaction to produce energy densified bio-coal," *Energy Convers. Manag.*, 88, 1086–1093, 2014. https://doi.org/10.1016/j.enconman.2014.04.071
[17] M. Wilk, A. Magdziarz, and I. Kalemba, "Characterisation of renewable fuels' torrefaction process with different instrumental techniques," *Energy*, 87, 259–269, 2015. https://doi.org/10.1016/j.energy.2015.04.073
[18] Y. Uemura, W. N. Omar, T. Tsutsui, S. B. Yusup, and S. Bt, "Torrefaction of oil palm wastes," *Fuel*, 90(8), 2585–2591, 2011. https://doi.org/10.1016/j.fuel.2011.03.021
[19] S. Matali, N. A. Rahman, S. S. Idris, N. Yaacob, and A. B. Alias, "Lignocellulosic Biomass Solid Fuel Properties Enhancement via Torrefaction," *Procedia Eng.*, 148, 671–678, 2016. https://doi.org/10.1016/j.proeng.2016.06.550
[20] K. M. Sabil, M. a. Aziz, B. Lal, and Y. Uemura, "Synthetic indicator on the severity of torrefaction of oil palm biomass residues through mass loss measurement," *Appl. Energy*, 111, 821–826, 2013. https://doi.org/10.1016/j.biombioe.2013.05.015
[21] S. Zhang, Q. Dong, L. Zhang, Y. Xiong, X. Liu, and S. Zhu, "Effects of water washing and torrefaction pretreatments on rice husk pyrolysis by microwave heating," *Bioresour. Technol.*, 193, 442–448, 2015. https://doi.org/10.1016/j.biortech.2015.06.142
[22] Y. Uemura, S. Saadon, N. Osman, N. Mansor, and K. Tanoue, "Torrefaction of oil palm kernel shell in the presence of oxygen and carbon dioxide," *Fuel*, 144, 171–179, 2015. https://doi.org/10.1016/j.fuel.2014.12.050
[23] D. A. Granados, R. A. Ruiz, L. Y. Vega, and F. Chejne, "Study of reactivity reduction in sugarcane bagasse as consequence of a torrefaction process," *Energy*, 139, 818–827, 2017. https://doi.org/10.1016/j.energy.2017.08.013
[24] C. Berrueco, D. Montané, B. Matas Güell, and G. del Alamo, "Effect of temperature and dolomite on tar formation during gasification of torrefied biomass in a pressurized fluidized bed," *Energy*, 66, 849–859, 2014. https://doi.org/10.1016/j.energy.2013.12.035
[25] M. Dudyński, J. C. van Dyk, K. Kwiatkowski, and M. Sosnowska, "Biomass gasification: Influence of torrefaction on syngas production and tar formation," *Fuel Process. Technol.*, 131, 203–212, 2015. https://doi.org/10.1016/j.fuproc.2014.11.018

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 171-185 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Complexity Drivers in Digitalized Work Systems: Implications for Cooperative Forms of Work

Benedikt Andrew Latos*, Markus Harlacher*, Florens Burgert, Verena Nitsch, Philipp Przybysz, Susanne Mütze-Niewöhner

Institute of Industrial Engineering and Ergonomics, Work Organization Department, RWTH Aachen University,

52062 Aachen, Germany

A R T I C L E I N F O

Article history:
Received: 01 August, 2018
Accepted: 16 September, 2018
Online: 27 September, 2018

Keywords:
complexity
digitalization
work organization
cooperative forms of work

A B S T R A C T

Digitalization changes the way people work to a considerable extent. It alters business models and process organizations of whole industries. The ensuing market dynamics and faster innovation cycles cause an increase in complexity. In this article, the interconnection of digitalization and complexity in work systems is analyzed. For this purpose, a framework for comparing relevant complexity definitions is developed. Moreover, complexity drivers in digitalized labor systems in six different organizational dimensions (process organization, organizational structure, technology, working conditions, product and personnel) are explored. 23 experts from the academic and industrial sector were interviewed using semi-structured interviews. The results of a qualitative content analysis show that the consideration of complexity and digitalization has extensive impact what becomes evident in interdependent relations amongst the organizational dimensions. Furthermore, complexity drivers in digitalized work systems are determined as a result of the analysis procedure. Finally, the implications of the expert interviews for cooperative forms of work are discussed. The concept of a "task complexity mountain range" is presented to explain the effect of task complexity on performance and motivation in the context of work groups.

1. Introduction

This article is an extension of work originally presented on the IEEE International Conference on Industrial Engineering & Engineering Management in Singapore (December 2017) [1].

Digitalization considerably affects work systems through diverse and quicker communication media as well as through the networking of manufacturing facilities and information systems. Industry 4.0, or the concept of advanced manufacturing, is used to manufacture individualized consumer goods or machines that are progressively equipped with mechatronic components [2]. Also, developments towards dynamic customer requirements change production planning for these products, since personalized customer requirements necessitate real-time and flexible reactions within a company [3]. This causes an increase in complexity in the process organization. According to [4], [5], though, the use of digital systems is able to reduce complexity in production planning, as they enable a more efficient and flexible control and management of company processes. However, the use of digital systems may cause interface problems or media discontinuities due to missing IT standards, since different IT systems are used for data acquisition that are not compatible with each other [6]. Yet, digitalization also enables the introduction of internal and external company network structures so that distributed and global cooperation across company boundaries is feasible [7]. Another significant change in the course of digitalization, which is discussed in literature, is the increase in the decision-making powers for employees that also partly results from flatter hierarchies [8], [9]. There is a general trend towards decentralized forms of management and control [10]. This implies that team and group work or interaction between individuals (in this article the term "cooperative forms of work" is employed) is relevant – especially in the course of digitalization. The changes result in an increase in the demands on employees. On the one hand, social competences are required. On the other hand, IT-related knowledge and skills in particular are becoming more important because digital technologies are used in almost all areas of companies and therefore must be handled by the employees [6], [9], [11], [12].

* Benedikt Andrew Latos, Markus Harlacher; Bergdriesch 27, 52062 Aachen, Germany, b.latos@iaw.rwth-aachen.de; m.harlacher@iaw.rwth-aachen.de

As all these facets might cause levels of higher complexity, it is important to install appropriate complexity management procedures in highly digitalized labor environments [13]. In order to manage complexity, it is necessary to understand the causes of complexity and its relationships, and to keep in mind that complexity may have both positive and negative effects.

This article investigates how the digital transformation relates to complexity in work environments. It examines work systems that use information and communication technologies which are interconnected to a large extent and are increasingly mobile as well as systems that process and supply significant proportions of information in digital forms (cf. [14]). Current complexity definitions from relevant research fields do not entirely incorporate important complexity characteristics with respect to a work system. Therefore, a framework for work systems is employed to systematize and analyze existing complexity definitions (section 2). Moreover, there is a lack of suitable classification systems for complexity drivers in digitalized work systems. In order to systematize complexity drivers that are explored with 23 semi-structured expert interviews (section 3), six organizational dimensions are chosen as a framework to present the results (section 4). Subsequently, complexity management approaches in digitalized work systems are outlined and digitalization-related effects on cooperative forms of work are discussed (section 5). The individual task complexity concept of Dalhöfer and Prieß [15] is extended to consider a work group. The article concludes with a critical discussion of the findings and outlines future research steps (section 6).

2. Complexity definitions from different research disciplines

The expression *complex* is widely employed in everyday language to express the difficulty of comprehending, investigating and forecasting relationships or systems' behavior that is not easy to understand. In research, the use and understanding of complexity depend on the specific research discipline [13], [16], [17], [18]. In order to conceptualize actions to handle complexity in digitalized work systems, it is vital to develop a theoretically grounded approach. Therefore, it is necessary to systematize existing definitions of complexity from relevant research disciplines in a framework that is suitable to represent a work system. As a basis for this framework the general model of a work system according to Schlick et al. [19] was chosen. This model describes the transformation process of input variables inside the work system into outputs. The working person that executes this transformation with work equipment is an essential part of the work system. Relevant research disciplines from the pluri- and interdisciplinary research field of industrial engineering and ergonomics that consider work system design are in particular: engineering, economics, sociology, and industrial and organizational psychology (see e.g. [19]). In order to develop an understanding of complexity for this framework, the understanding of complexity within these research disciplines is examined below:

Although diverse scientific disciplines employ the expression *complexity* in different manners, general system theory is widely used to define the phenomenon of complexity [16]. Since a work system can be regarded as a general system, the characteristics of complex systems are outlined: systems that are meant to be complex have lots of system elements as well as they exhibit many interactions simultaneously [20][21]. Relevant literature on system theory indicates that certain characteristics of complex systems are the number and type of elements and interactions, rate of change of the system structure and uncertainty of the current and future system status (various complexity definitions can be found e.g. in [16][22]). In fact, von Bertalanffy [23] states that the characteristics of a system are due to the properties of its elements and the connections amongst them. The construct of emergence is special for complex systems. This phenomenon means that the system's behavior is not entirely explainable with the properties of the single system elements [24]. Schuh and Schwenk [25] distinguish between simple, complicated, relatively complex and extremely complex systems, based on their number and diversity of the elements and relations as well as their variability and dynamic (see Figure 1). Whilst simple systems are characterized by few elements, interrelations and behavioral possibilities, extremely complex systems have a multitude of elements with a great variety of interrelations and possible system behaviors with variable impacts of the elements [26].

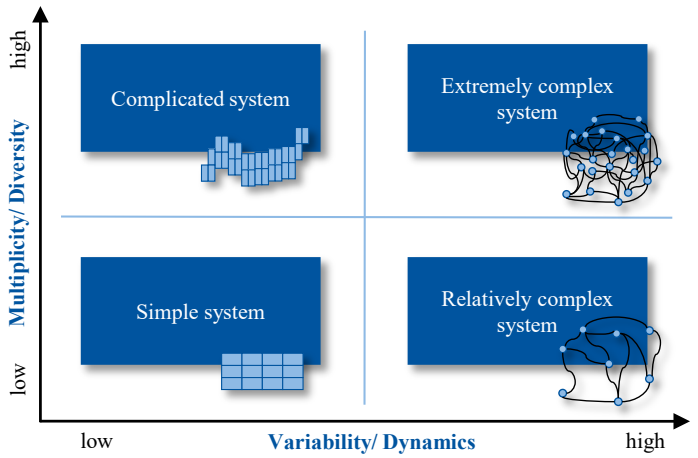


Figure 1: System states according to Schuh and Schwenk [25], Ulrich and Probst [26]

System theory is also frequently used in engineering disciplines to explain the phenomenon of complexity. Therefore, the complexity properties from general system theory also apply to considerations of complexity in an engineering context [27], [28], [29]. For example, system elements in a labor system are produced goods, tools, assignments or personnel. The system's interrelations are, for instance, relationships of the elements during production or communication procedures. Because digitalization in work systems could have an impact on the elements or relationships in the system, this could also change the level of complexity in it.

In the literature, researchers distinguish static or structural complexity from dynamic or operational complexity when manufacturing systems are regarded [27]. The two facets are often generalized as objective complexity, whereas Blockus [22] describes complexity in the field of business administration with reference to system theory and includes perception as a subjective complexity characteristic. This aspect is crucial for a work system framework, since workers operate in these systems. For production systems, subjective complexity reflects how individual persons

experience complexity, e.g. with respect to the design of single work stations in assembly [30]. Zeltzer et al. [31] state that complexity perception in assembly is largely determined by the operator and therefore define complexity as subjectively negative perceived effects of the assembly system on the operator.

Persons in work systems have to make numerous decisions whilst performing their jobs. This is why the complexity understanding in psychology is important for the framework. Within the discipline of psychology, complex situations and problems are described by interrelated variables, a considerable level of dynamics, a lack of transparency concerning a situation and multiple goals that are conflicting (polytely) [32][33][34]. Willke [35] provides a sociological definition that considers the degree of multidimensionality, interconnectedness and relevance of the consequences of a decision problem. Complex systems have diverse options that can be chosen so that operators have to select an alternative [36].

The aforementioned complexity definitions contain central characteristics, which are relevant for a complexity definition with regard to a work system. For a systematic comparative analysis of existing complexity definitions, which relate to the chosen framework, relevant criteria are required:

In this article, work systems are considered with regard to the transformation of work through digitalization. Therefore, a system-theoretic criterion "system size" should be defined that considers the type and number of elements, as well as interactions between them. As second criterion, "dynamics and uncertainty" should be included. This criterion considers the non-static and emergent behavior of complex systems. Concerning the work system framework, the system-theoretical concept should be expanded to consider humans and effects of working conditions on humans. Decision behavior of human operators in complex situations within their work systems should be taken into account. Therefore, "polytely" is defined as a criterion for complex situations with conflicting goals. In addition, complexity may arise due to human information processing and differing semantic perceptions. Following the stress-strain-concept [37][38] and the individual perception of complexity according to Blockus [22], it has to be stated that human perception of complexity is subjective. Thus, the criterion "subjectivity" is introduced. With respect to the transformation of work through digitalization, the aspect "digitalization" is explicitly included. Figure 2 presents a selection of complexity definitions from relevant research disciplines with regard to the derived criteria.

Author(s)	Research Discipline	Definition of Complexity/ Complexity Characteristics	System Size	Dynamics and Uncertainty	Polytely	Subjectivity	Digitalization
von Bertalanffy (1972)	System theory	Number of elements, variety of elements, number of interrelations	●	●	○	○	○
Ulrich and Probst (1988)	System theory	Complex systems are characterized by a high number and diversity of the elements and relations as well as high variability and dynamics. This allows complex systems to be distinguished from simple and complicated systems.	●	●	○	○	○
Deshmukh et al. (1998)	Engineering	Static complexity of a production system is characterized by its structure, variety of subsystems and strength of the interactions.	●	○	○	○	○
Sivadasan et al. (2006)	Engineering	The dynamic or operational complexity is defined as uncertainty regarding the temporal change of the system.	○	●	○	○	○
Blockus (2010)	Business Administration	The main characteristics are the number and type of elements and connections and the variability over time. Beside these objective complexity characteristics, perception is considered as a subjective component.	●	●	○	●	○
Zeltzer et al. (2013)	Engineering (assembly system design)	Complexity of a workstation is the sum of all technical and ergonomical aspects and factors that make the set of tasks to be performed within it by an operator mentally difficult, error-prone, requiring thinking and vigilance, and inducing stress.	○	○	○	●	○
Dörner et al. (1983), Dörner (2011), Kluge (2004)	Psychology	A complex problem is described by four characteristics: dependencies between variables, internal dynamics, non-transparency of the situation, multiple conflicting goals (polytely).	◑	●	●	○	○
Willke (1996)	Sociology	Complexity is described by three characteristics: multidimensionality, interconnectedness and relevance of the consequences of a decision problem.	◑	○	◑	◑	○
Luhmann (1991)	Sociology	A system is complicated with a large number of elements and connections, whereby a complex system simultaneously contains many alternative options and forces selection.	●	○	●	◑	○

Explanation for evaluation criteria: ● fulfilled; ◑ partially fulfilled; ○ not fulfilled

The results illustrate that none of the considered definitions fully covers all defined criteria. Striving towards a definition containing all criteria, a synthesis of the discipline-specific definitions is required. Additionally, none of the definitions includes the aspect of digitalization. As a conclusion, the findings can be used to derive an integrated comprehension of complexity characteristics in the context of digitalized work systems:

- Complexity in a system refers to a construct of higher order that is composed of single system elements.
- A complex system, as sum of its elements, has the properties of high dynamics caused by multidimensional interactions, unpredictability of the system state and polytely, inducing self-generated characteristics of the system.
- Concerning the digitalization of work systems, objective complexity has to be distinguished from subjective complexity. While subjective complexity is induced by information processing and perception of the individual, objective complexity arises from interactions of internal system relations and external environmental impacts.

Especially from a work design point of view, it is crucial to develop a deep comprehension of variables that induce complexity in a digitalized work environment. On the one hand, labor systems with a too high degree of complexity are avoidable. On the other hand, operators may be qualified to manage complexity, if they are provided with sufficient information via support systems within work procedures. Moreover, complexity is also mentioned in terms of knowledge management as a basis for developing flexible and self-learning organizational structures [39]. All these aspects can only be considered, if there is a fundamental knowledge of complexity drivers in a work system. In this article, the understanding of the term "complexity driver" is based on the general definition of Vogel and Lasch [40], which they developed from the use of the term in the literature:

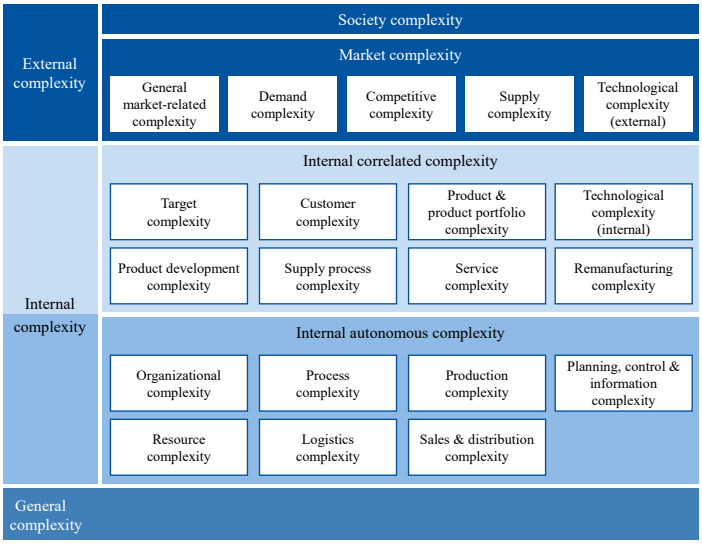
"Complexity drivers are factors, which influence a system's complexity and company's target achievement. They are responsible for increasing system's *[sic]* complexity level and help to define the characteristics or the phenomenon of a system's complexity. Complexity drivers are influenced by one another, that is by internal or external drivers, and cannot be reduced completely to another one" ([40], p. 18).

As another result of their systematic literature research, Vogel and Lasch [40] show that there is a multitude of different approaches to identify and classify complexity drivers. They identify complexity drivers with a systematic literature research procedure and employ a quantitative frequency analysis of mentioned complexity drivers. They present a general complexity driver classification system, based on existing classification systems, which is shown in Figure 3.

External complexity drivers are distinguished from internal and general complexity. The internal complexity can be actively influenced by the company. It is further divided into correlated complexity drivers, which are directly related to external complexity impacts, and autonomous complexity drivers, which are not directly influenced by external impacts. The general complexity category includes all drivers that cannot directly be assigned to other specific categories (e.g. uncertainty). The rectangles in Figure 3 represent one complexity driver category that is further subdivided into numerous complexity drivers. Detailed descriptions of the assigned complexity drivers are presented in [40].

However, it can be seen that digitalization is not explicitly included as a complexity driving factor. This also applies to alternative complexity driver systematizations (see e.g. [41][42][43][44]). Furthermore, diverse methods exist in order to assess the level of complexity in work systems (see e.g. [41][42][45]). However, these approaches also do not explicitly take changes into account that are caused through digitalization.

Above all, these classification systems regard complexity on a general or company-wide level. They do not focus on complexity drivers that directly relate to work of the persons in a company and their specific work systems. This necessitates that a suitable framework has to be used to explore drivers of complexity in digitalized work systems. This framework is presented in the next section.



[40]

3. Methodology

As a framework to systematize drivers of complexity in digitalized work systems, six organizational dimensions were chosen. The organizational dimensions were the process organization, organizational structure, technology, working conditions, product and personnel. The authors have employed this framework already in diverse transformation projects in different industry sectors. In these projects, the dimensions have proven to be applicable for holistic organizational transformation projects regardless of industry settings.

To identify drivers of complexity for every organizational dimension in digitalized work systems, semi-structured expert interviews were performed. This methodology was selected, because it provides valuable qualitative insights at an initial exploration stage. Moreover, it can be used to complement ongoing quantitative research steps. The questions consisted of a constant number of open questions. This procedure guaranteed that

all important questions were treated in all interview sessions. However, the interview design made it possible to stress on aspects that are especially important for an individual interviewee at the same time. Also, the aspects could be discussed according to the individual reference framework. The duration of the interviews was about 45 minutes. The interviews were transcribed according to Mayring [46] by using the audio recordings with permission of the interviewees. In one case, no recording was available so that interview notes were used. The interview transcription time took up to 8 hours for one interview. In total, 23 interviews were performed. 22 interviews were conducted via phone. One was a personal interview. The majority was conducted as phone interviews because of interview economics reasons: the interviewees worked in companies that are spread throughout Germany. Figure 4 gives an overview on the experts and their backgrounds that were included in the study. The interviewees were acquired via using the databases and business networks of the research institute. They were requested via e-mail. In total, about 30 interview requests were sent, whereof 23 interviewees accepted the invitation. For being taken into account as an expert, two requirements had to be fulfilled: the experts needed to have privileged access to information and must have been responsible for conceptualizing, implementing or controlling problem solutions with regard to digital strategies [47].

The 23 interviewees held positions in German academia or industry and were requested because of their positions and knowledge of digitalization and complexity management or their expertise of work organization. Moreover, the experts belonged to industries from both production and service. Most worked in the fields of engineering, the management of projects, consignment and logistics, information and communication technology, banking or within the media industry.

Also, the experts were representatives of entrepreneurs or belonged to committees that represent employees. To prevent biased analysis results, the expert acquisition procedure was conducted in a manner to obtain interviewees that cover several domains and stakeholder groups.

Number	Sectors / Domains	Function of the expert	Comment	Number of employees
1	System house	Director of system house	Development and consulting of IT system houses (focus on process organization)	< 50
2	Science	Professor information management	Focus on hybrid services, digital business models	< 50
3	White goods	Head of Department assembly	Cycled assembly system (large quantity)	> 10,000
4	Software company	Customer adviser product lifecycle management	Cloud computing, Big Data analysis	> 1,000
5	White goods	Head of Department industrial engineering	Digital assistance systems with work instructions, error feedback via apps	> 10,000
6	Automotive	Expert industry 4.0 in corporate work council	Focus on digital work, industry 4.0	> 100,000
7	Start-up electro mobility	Chief information officer	Agile, digital product development, smart products in the automotive sector	< 250
8	Science	Director of demonstration factory	Demonstration Factory Industry 4.0	< 250
9	Science	Expert industry 4.0	Holistic production and enterprise system design, project management	< 250
10	Logistics, consignment	Chief executive officer	System solutions for working with smart glasses	< 50
11	Industrial Connectivity	Head of Department global digitalization and intelligence	Data connection of world-wide facilities	> 1,000
12	Automotive	Complexity manager	Focus on the use of digital analysis methods	> 50,000
13	Service, maintenance and repair	Key account manager	Proactive maintenance through the use of digital tools	> 1,000
14	Consulting and research	Director of consulting company	Enterprise organization	< 50
15	Media	Chairperson of work council	Print media	> 1,000
16	Media	Chairperson of work council	Online media	> 250
17	Information and communication technology	Chairperson of work council and member of the supervisory board	Technology and Innovation Committee	> 10,000
18	Logistics	Deputy chairperson of corporate work council	Member of the IT Committee	> 100,000
19	Banking	Member of work council	Member of committees for the adoption of resolutions, e.g. on topics related to digitization	> 1,000
20	Banking	Quality manager and member of work council	Payment service provider	> 10,000
21	Logistics, consignment	IT-expert	System administrator	> 50,000
22	Information and communication technology	Director Human Resources	Operation of IT systems for customers	> 10,000
23	Academia	Director of university campus	Research and teaching on industry 4.0	< 250

The questionnaire guide was divided into four parts. In a first step, demographic information on the interviewees was documented. The second questions asked for aspects that change companies and labor through digitalization; e.g. question were: what are the main digitalization-related changes in your company within the last 10 years? Which important changes do you anticipate for the next 10 years? The third question part focused on whether and how digitalized labor systems result in higher complexity levels. Therefore, the experts should name complexity drivers which are caused by digitalization for the six organizational dimensions respectively. In the final part of the interview, the experts were asked to identify measures to manage complexity in digitalized labor environments. Figure 5 summarizes the content of the interviews. The detailed questionnaire guide is included in Appendix 1.

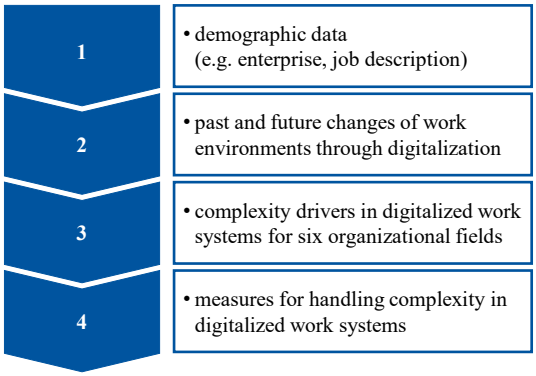


Figure 5: Structure of the interview template

The analysis process initially incorporated a verification of the transcription documents. This was conducted with the available sound recordings of the interviews. Then, the single interviews were qualitatively analyzed via employing the methodology of qualitative content analysis [48][49][50].

A qualitative content analysis is a systematic and intersubjective verifiable text analysis procedure that meets the needs of interpretation and semantic meanings of a linguistical research basis [51]. In contrast to quantitative content analyses, the main purpose is not to discover quantitative connections, but rather to reduce a large amount of analysis material and to extract the main contents. This makes the method especially useful for explorative studies. Mayring [48] describes a systematic and rule-based analysis approach. The structured and transparent procedure enhances the reliability and validity of the analysis. In this case, the general procedure was adapted to the specific research aim.

In a first step, a qualitative-interpretative analysis of every interview was conducted. As code unit, a single word was chosen and a whole sentence structure was defined as context unit. Furthermore, the analysis unit was structured according to the interview template. Within the analysis procedure, the statements of the experts were generalized, paraphrased and reduced to elementary categories. The categories were derived by using a hybrid approach of deductive and inductive category formation. The inductive approach was chosen, since the research did not aim at confirming existing theories. The aim was rather to extract general statements from the interview material. However, deductive analysis elements were included, as statements were assigned to expressions that are well established in literature. For example, the expression "polytely" (cf. section 2) was chosen as category name for expressions that contained phrases such as "conflicting goals" or "different goals that cannot all be reached at the same time". This simplified a purposeful assignment of the interview phrases.

In a second step, the single analysis results were combined and aggregated with a quantitative analysis of mentioned categories. This made it possible to describe how often a certain category was mentioned by the interviewed persons. By doing so, complexity drivers were extracted via counting the number of mentions for single driver categories within all 6 organizational dimensions. Figure 6 summarizes the whole analysis procedure.

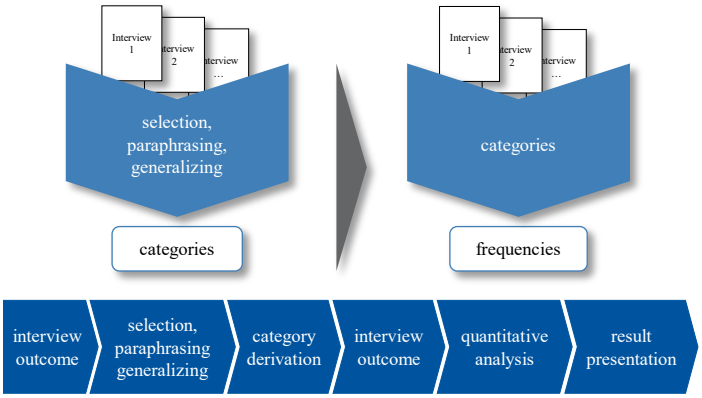


Figure 6: Procedure of qualitative content analysis according to Petz et al. [52], Duckwitz et al. [53]

According to Mayring [48], a qualitative analysis should meet two major quality criteria. On the one hand, intra-code-reliability requires that a repeated analysis procedure of the same coder after a certain time should lead to similar results. On the other hand, inter-code-reliability means that different persons will achieve similar analysis results. As the method is an interpretative approach, there will always be slight differences in the results, e.g. concerning the wording of categories between different coders (see e.g. [48]). With this analysis procedure, one repeated analysis by the initial coder as well as one further analysis with an additional coder was conducted. In both cases, similar results were extracted from the interviews which can be seen as indication for meeting the reliability criteria.

4. Results

In this section, the results of the aforementioned qualitative content analysis procedure are presented. In the first subsection, past and future changes of work environments through digitalization that were extracted from the interviews are outlined. Subsequently, the derived complexity drivers are presented for every organizational dimension, respectively. In the third subsection, the results for measures to handle complexity in digitalized labor systems are displayed.

4.1. Past and future changes of work environments through digitalization

Concerning the last 10 years, the experts pointed out diverse changes within their companies that have been caused through digitalization. The main changes are due to the new opportunities of manufacturing customized products. In addition, a development

towards hybrid products in production industry was named. This means that the product is enhanced with further service options for the customer.

Also, the experts stated that former manually documented procedures have been changed into digital processes. These workflows incorporate several companies so that the dynamics and speed of these procedures have considerably increased as well as the duration for reactions has become shorter. Enterprise software, information and communication tools and cheaper sensor technologies offer the opportunity for companies to track and process big amounts of data. However, the operators within the companies very often have a problem to interpret and mentally capture these masses of information. As a solution for this problem, the use of big data algorithms was proposed.

Employing information and communication technology remarkably altered every day work procedures, since the amounts of interpersonal face-to-face communication are reduced. Nowadays, (digital) work forms of collaboration can be observed increasingly. Portable devices such as mobile phones, tablet computers and netbooks make it possible to perform the job in flexible work forms in terms of location and time. This gives employees more flexibility and enhances family friendliness of enterprises, whereas the persons at the same time have the drawback of possibly being always available. The interviewees concluded that these changes rather relate to administrative departments and service sectors than to manufacturing. Within the service sector, value creation networks have a higher level of digitalization than in manufacturing.

With respect to the next 10 years, the interviewees predicted increasing and quicker rates of process changes. Consequently, data will not be documented manually any more. Planning procedures will be complemented with digital support tools as well as data in real-time will be provided. Concerning manufacturing steps, the interviewees pointed out that planning and controlling would change from an internal enterprise-focused view towards an approach that considers whole value-chains. Furthermore, business models will more and more tend to include services. Therefore, products will include more software and will be a medium to offer services to the users of the products. For instance, a car manufacturer will no longer be only selling cars but rather mobility, since the car offers supplemental services to the customer.

Although experts that represented employees feared big job losses because of digitalization, the majority of the interviewees agreed that operators would not be replaced in labor systems. However, different tasks and qualification requirements will develop so that there will be new job types.

4.2. Complexity drivers in digitalized work systems

Although the interviewees did not all have the same occupational backgrounds, the overall analysis indicates that similar points can be generalized from the single interviews. In total, the analysis results show that a considerable development towards very adaptive procedural and structural organizations is required for meeting the environmental demands and to evolve to customer-focused production procedures. This development is made possible through digitalization of manufacturing and service procedures. Nevertheless, this causes a higher complexity level.

Regarding the organizational dimension of process organization, the experts answered that complexity increases due to more customer involvement into faster work procedures. This calls for reactive and adaptive planning processes. Since digitalization can be used to initiate reactions in a company in real-time via using novel technology, this remarkably raises the dynamics of the whole organization. The change towards worldwide active value creation networks with several levels also increases the overall complexity. Since labor systems are more digitalized, the amount of human machine interfaces and diverse system solution rises. In this context, parallel IT-systems may increase the complexity level because of several maintenance and data processing steps. This may in turn cause data redundancies and problems of lacking data hygiene. Furthermore, media changes during work processes were mentioned as roots of complexity, which is digitally solvable via solutions for integrated systems. Figure 7 summarizes the drivers in the dimension of process organization and their number of mentions. Since every expert could mention several or no complexity drivers for every organizational dimension, the number of mentions does not necessarily have to sum up to the number of experts.

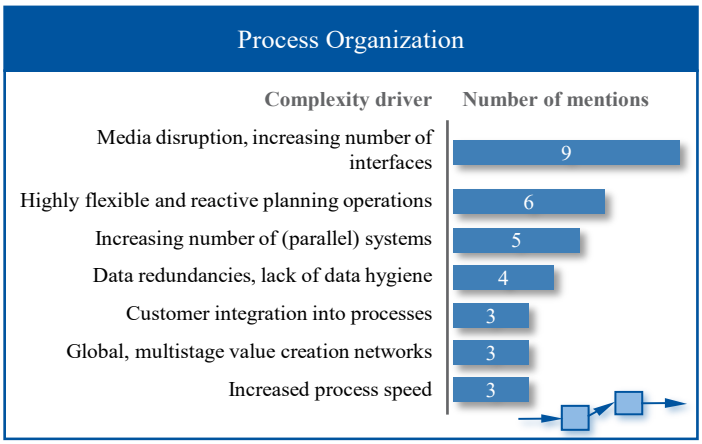


Figure 7: Complexity drivers in digitalized work environments concerning process organization, including their number of mentions

To develop such adaptive enterprise organizations, a shift of traditional hierarchic organizations towards the direction of flatter hierarchies is necessary. Almost all interviewees agreed on the point that agile work methods and project work would be most frequently found in the course of digitalization. In this context, complexity with respect to the organization increases, when a development towards trans-disciplinary work in (worldwide) structures is needed. This causes more coordination procedures. To work in smaller and agile cooperative structures also causes that employees need to be given more autonomy. The targeted final state may make it easier to fulfill market demands. However, the procedure of changing the culture in an organization was seen as raising the complexity level in a first step. If work persons are not empowered and qualified in diverse dimensions, occurrences with conflicting goals (polytely) can imply complexity for single operators. Complexity drivers and their number of mentions regarding the dimension of structural organization are illustrated in Figure 8.

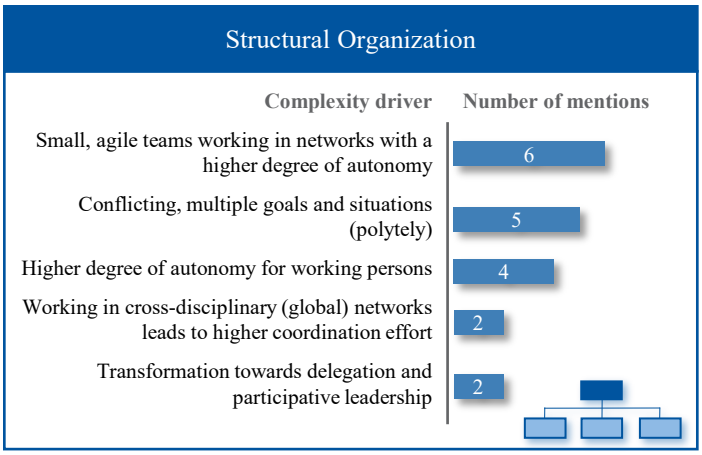


Figure 8: Complexity drivers in digitalized work environments concerning structural organization, including their number of mentions

For the organizational dimension of technology, the experts named rising requirements for IT security as most important aspect in digitalized labor systems. Systems that are not compatible, missing technological standards and lacking common sets of protocols increase complexity during digitalization. These aspects gain importance, since harmonization problems happen because of systems that have evolved over a long time span. Furthermore, much higher information amounts need to be captured in real-time with software to analyze the data. Faster innovation cycles with respect to software and hardware cause more dynamics of introducing novel systems. Figure 9 summarizes the drivers concerning technological aspects and their number of mentions.

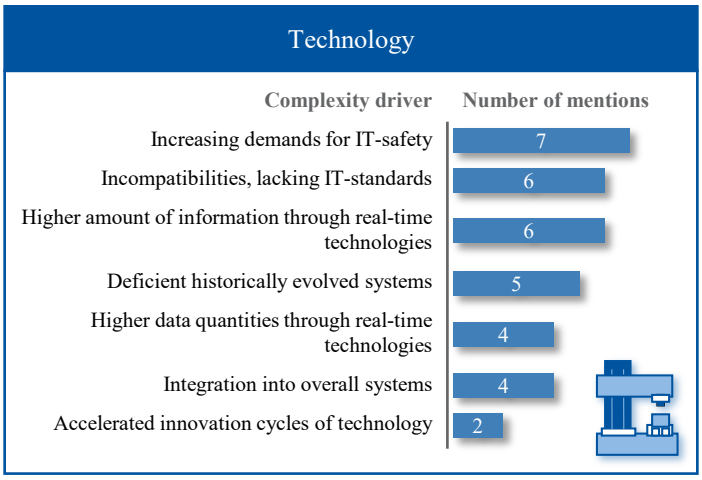


Figure 9: Complexity drivers in digitalized work environments concerning the technology, including their number of mentions

Digitalization will also have an effect on the conditions of work and will make new methods of digital communication possible. However, mentioned by the interviewees, only communicating digitally could not totally replace interpersonal communication and cause communicational complexity, because the semantic meaning of expressed phrases is mainly strengthened with face expressions or gestures. For this reason, hybrid communication ways should be considered. The majority of the interviewees stated that the digital transformation in general changed jobs into the direction of tasks that have to manage more complexity. Although monotonous tasks are likely to be replaced by automation, jobs that demand analytical information processing, creative thinking and the capability of solving problems combined with specialist skills are more and more required. This causes that employees need transdisciplinary abilities (for instance, qualifications such as self-organization or multidimensional work with respect to diverse disciplines). Task complexity occurs as well, if novel digital systems are used and if more digitalized information has to be captured mentally in a shorter time. Lots of interviewees stressed on the point that this facet already produced trouble for employees in many cases. If the rising amount of operator machine interfaces was not created in a way that they can be used easily, a higher level of complexity would be a consequence for the employees. Nevertheless, single interviewees mentioned that the use of digitalization could at the same time be employed to make work easier. For example, work instructions can be provided during the task fulfillment. A further point is that the digital transformation could cause more task quantities. In this context, scan devices facilitate to introduce inventory documentation jobs into work procedures. All in all, a major complexity cause in digitalized labor systems was identified in the point that digitalization frequently is not very well established in management hierarchies. Particularly for worldwide operating organizations, the significance of different laws and agreements for data protection in diverse nations has extensive consequences. Figure 10 illustrates complexity drivers in the dimension of working conditions.

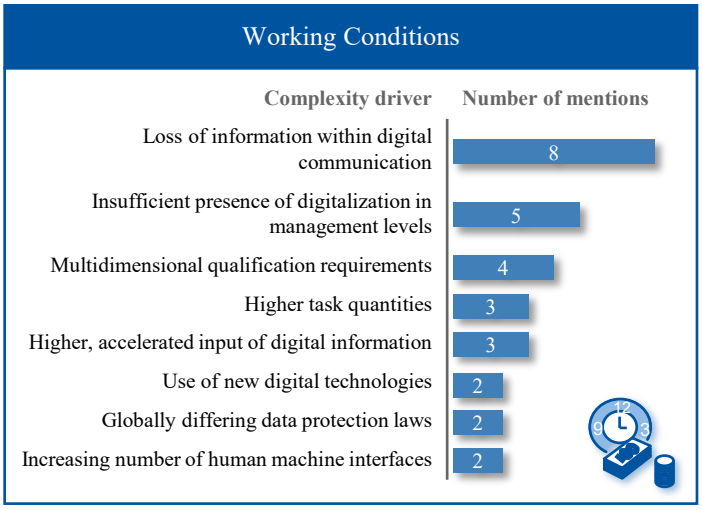


Figure 10: Complexity drivers in digitalized work environments concerning working conditions, including their number of mentions

With respect to the organizational dimension of products, the interviewees named individualized products and more product variants as main complexity drivers. In general, digitalization causes product and service portfolio complexity because sensor devices, software and connecting elements are important parts of a product. This implies that designers of goods and machines will have to take into account that the final product is going to be used in a web-based context. These changes remarkably enhance feasible functions of the product but also cause more software dependency. The radical development from typical products towards hybrid production necessitates that the whole life-cycle of a produced good will need be regarded during the development of the product. Since producers and customers will not only interact once when selling a smart product, they will likely be in close contact during the whole life-cycle when processing data packages

are transferred. This provides numerous novel options and makes a product more complex. For instance, individualized services or upgrades may be suggested as well as novel business models that are based on data analyses will come into focus.

In addition, very iterative and accelerated product development procedures in which clients work with the contractor together were mentioned as complexity drivers. This process of co-development between companies is essential, since a considerable gain in IT-skills and system thinking develops on the client side. In particular within the service sector, digital opportunities are employed to integrate the customer (business-to-customer relation) into the development of new services and into the service delivery. In the context of self-service, digital procedures make it possible to outsource tasks to the customers and to let them directly take part in the service fulfillment. In analogy to the organizational dimension of technology, the interviewees mentioned that digitalization fastens the release cycles for new innovations or products so that the complexity level of the products increases. Figure 11 summarizes the complexity drivers concerning the product.

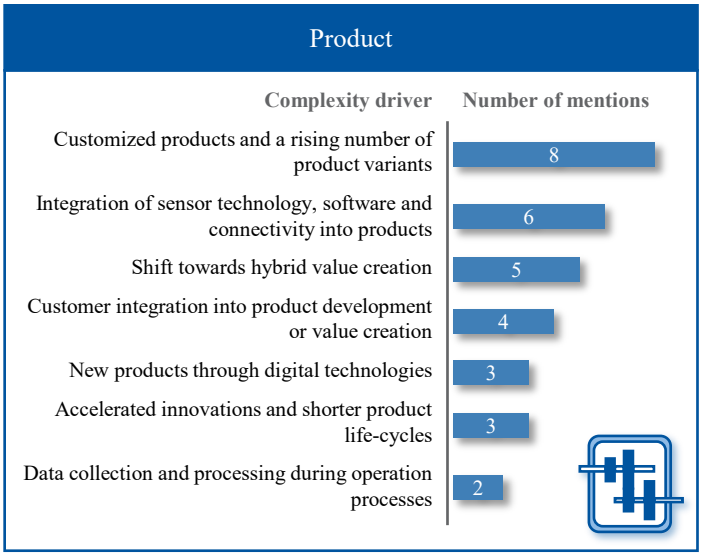


Figure 11: Complexity drivers in digitalized work environments concerning the product, including their number of mentions

Concerning the organizational dimension of personnel, the interviewees said that qualification requirements for more digital understanding and skills are the most important aspect for handling the complexity in digitalized labor systems. However, the operators very frequently do not have the needed competencies. Moreover, a major gap of IT understanding between employees can be detected. Also, digitalization offers the possibility to work in mobile and flexible work conditions. This allows for balancing the private life and the job. Nevertheless, this also can cause very complex situations for the staff, because contrary requirements could evolve from the trend that makes the separation between the job and free time more difficult. The employment of mobile gadgets also implies that an employee is always available. This can be the cause of a dependency on technology to some extent. Additionally, severe problems were mentioned that concern the aspects measuring and controlling the performance of the personnel in labor environments. If new places of work are conceptualized for digital work systems, the complexity of planning increases due to missing approaches for strain measurement. Furthermore, there is a lack of rules and relief concepts for digital work. Ultimately, many experts stated that digitalization offers novel job forms, such as crowdworking. This may have major implications for single persons or whole organizations. In addition, this can also imply complex problems for societies. These could be related to topics of social exclusion or regulatory problems for whole systems due to lacking contributions to social insurances. Figure 12 summarizes the complexity drivers concerning the personnel.

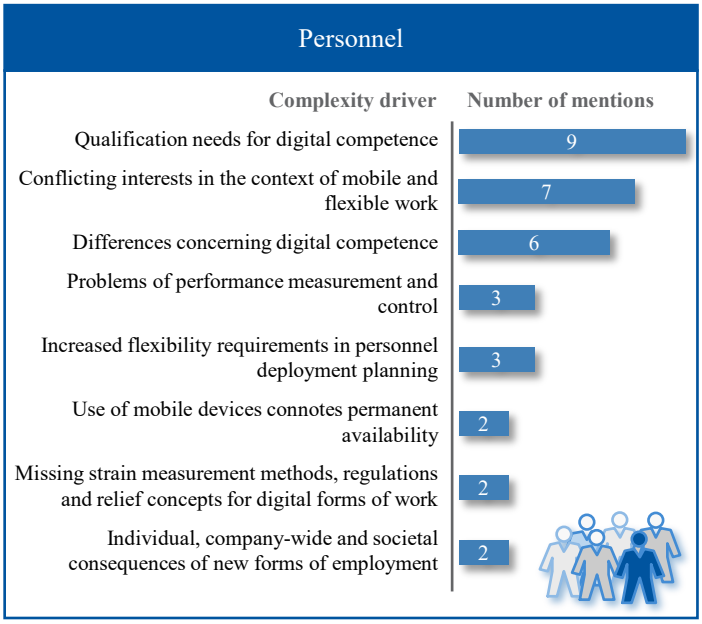


Figure 12: Complexity drivers in digitalized work environments concerning the personnel, including their number of mentions

4.3. Measures to handle complexity in digitalized work systems

The interviews were also used to explore measures to manage complexity in digitalized work systems. Adequate training of employees was most frequently mentioned by the experts. On the one hand, this implies qualifications for imparting methodical and expert knowledge on new technologies, processes and work equipment. On the other hand, training includes enhancement of social and communicative competences, which are targeted e.g. at avoiding or solving communication problems and conflicts in cooperative decision-making processes. Training was also considered as necessary to support employees coping with mental and physical strain as a consequence of changing or increasing job requirements.

The second most frequently mentioned measure was participation. Employees that are affected by digitalization-induced change should not only receive adequate training, but also should be involved in change-related development and decision-making processes at an early stage. Participation in pilot projects for implementation of new technologies and communication tools was mentioned as an example in order to enhance acceptance for digitalization-related innovations and to ensure economic efficiency. Additionally, participation-oriented pilot projects were recommended to design user-centered support systems and identify policies or required formalities on the following aspects: workload, working time, occupational safety and data protection.

In this context, work in agile teams was, for example, seen as form of work that especially requires policies concerning these aspects.

Development and use of algorithms for automatic data analyses were proposed as further measures to handle complexity. Thus, increasing data volume and diversity in the course of digitalization could be limited to the information and key figures required for the specific application. These and similar measures aim to relieve employees from analyzing and processing large amounts of data. The resources saved can be used for the accurate interpretations of data, which still need to be carried out by employees.

To harmonize different systems and their interfaces, the development and introduction of reference architectures for IT systems were mentioned. A modular structure of IT systems enables further development and easier linking of new technologies to the overall IT system. The definition of IT standards and the standardization of the IT structure help to avoid interface problems and to ensure the interoperability of the systems used. According to the experts, simple operation of products and systems is an equally important measure, e.g. through intuitive user guidance and ergonomic interface design. While this results in potentially more demanding programming, there are various benefits: avoidance of handling problems and errors, reduction of operating effort, and reduction of qualification effort during introduction and application of newly introduced technologies and systems.

The experts also mentioned the development of more transparent business processes to clarify responsibilities and avoid misunderstandings in work instructions due to inaccurate communication or misleading information. This also aims at a solid understanding of processes of all employees and managers as well as better overall communication.

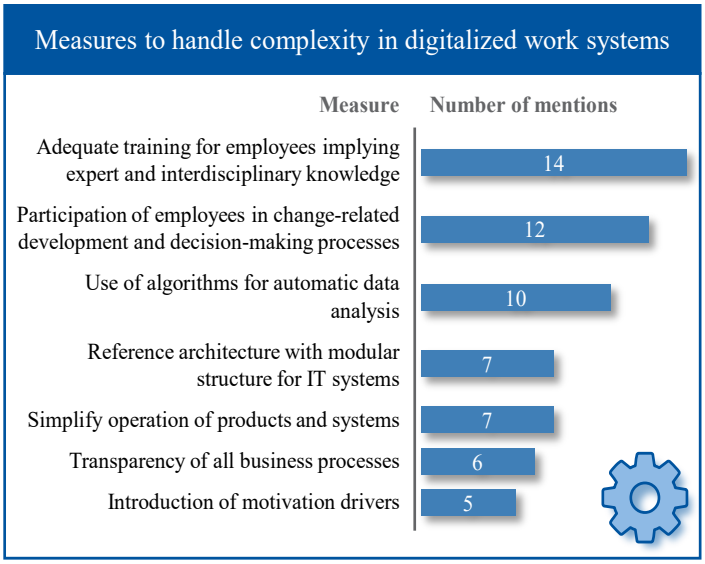


Figure 13: Measures to handle complexity in digitalized work environments including their number of mentions

Finally, the introduction of motivation drivers includes measures that can or should contribute indirectly to mastering the resulting complexity by promoting the willingness of employees to change. Motivation drivers could be e.g. suitable break concepts, more varied activities or "gamification" approaches combined with monetary incentives. Figure 13 illustrates the categorized measures to manage complexity in digitalized work systems as a result of the interview analysis.

5. Implications for cooperative forms of work

The results of the expert interviews have far-reaching implications for cooperative forms of work. In this context, the term cooperative form of work is employed to express that people work in a team or group structure and thus exhibit interaction amongst each other whilst performing their job. This includes a project team as well as a work group in a production environment.

In the literature, there is already extensive research on team or group work that describes the interrelations of relevant variables, e.g. in team models (e.g. [54][55][56][57]; overviews can be found in [19][58][59]). These models often focus on teamwork in the context of innovation and do not consider complexity parameters. Only a few conceptual and non-empirically validated models (e.g. [60]) contain "task complexity" in the context of cooperative forms of work as a single complexity-related variable. The expert interviews revealed that the ratio of cooperative forms of work is expected to grow in the field of production. A literature review of existing models for cooperative forms of work shows that currently there is no interrelated and empirically assessed model for group work in production and thus no consideration of complexity in a superordinate model (see e.g. [59]). Therefore, further research is necessary in this field. Such a model could be developed conceptually as an input-process-output model (see e.g. [19]). This approach enables to investigate the effects of digitalization and complexity as input factors on process and output dimensions. The interview results offer diverse aspects that directly refer to cooperative work. These aspects could be integrated into the development of such a new model. By doing so, the explorative results of the interviews, which originally did not aim at confirming existing theories, could be used to develop hypotheses, new theories and an overall framework for analyzing cooperative forms of work.

When regarding the dimension of structural organization in the interviews, a transformation towards cooperative forms of work was the most frequently mentioned complexity driver for digitalized work environments. Small and agile teams working in network structures with a high degree of autonomy were considered by the experts as a predominant form of work in the late stage of digital transformation processes. This can be seen as a continuation of the trend observed by the experts, which already describes a change towards cooperative forms of work over the past 10 years. These results strengthen that team and group research will increasingly come into focus – particularly in the course of digitalization. This conclusion is also reached by Trompisch [61].

Hence, there is a need for further research on group processes on the one hand and to explicitly focus on the effects of digitalization on cooperative forms of work on the other hand. In addition, the results of the expert interviews showed that digitalization does effectively imply complexity-driving aspects that have an impact on working persons. For this reason, it is also necessary to apply complexity considerations to cooperative forms of work and to investigate complexity in greater depth in this context.

Assuming that humans do not have a stable input-output relationship [62] supports the consideration of human behavior and perception in complexity studies, e.g. with regard to emergent events [63]. For example, a user of a new digital system could use the system for a completely different purpose compared to the original goal of the system. If this individual view was transferred to a work group consisting of several individuals, it seems trivial that the interaction of several people can especially be classified as complex. From a system theory point of view, a work group, e.g. a team of workers in production, can be seen as a complex structure or subsystem within the complex company system. In this context, the characteristics of complexity from system theory that were derived in the first section of this article can be considered and transferred to the system "work group".

On the whole, it is essential to consider factors such as digitalization and complexity in a differentiated manner. Complexity may contain positive and negative aspects concerning both organizational levels (e.g. [17] [64]) and individual levels. On the company level for instance, a larger variety within a product portfolio – and thus greater complexity – can be used to respond to changing market requirements and thereby to achieve competitive advantages. However, an excessive broad product range also bears the risk of strong cross-subsidization effects and thus economic losses (see e.g. [65] [66]). As the focus of this article lies on work design, the individual level concerning work persons will be discussed in more detail.

For an individual person and in analogy to the Yerkes-Dodson law, Dalhöfer and Prieß [15] conceptually describe the influence of task complexity on the individual performance and motivation in the context of indirect business processes with an inverted u-shape parabola. The decline in performance for tasks that are less complex than the optimum task complexity is explained with demotivation due to highly monotonous activities. The area near the parabola's apex describes an optimum of task complexity in which the worker is challenged in a positive sense by the work task and thereby is highly motivated. The decline in performance to the right of the complexity optimum is caused by the employee being overloaded by the work task. Regarding the individual skills of the working person, the task becomes too complex so that work performance and motivation decrease. Since complexity is not clearly measurable due to the subjectivity of its perception, and a slight change in complexity does not necessarily lead to immense changes in performance, an interval for acceptable task complexity should be defined. Dalhöfer and Prieß [15] use the term "Eu-Stress" (good stress) to define the term "Eu-complexity" (good complexity). This term describes the range around the complexity optimum in which a high performance is achieved. In analogy, they describe the area of overload as "Dys-complexity". Figure 14 illustrates these conceptual considerations.

It is important to note that the curve trace of the parabola and therefore also the interval boundaries vary in analogy to the stress-strain concept [37][38] between different persons.

In general, the individual parabolic curve can be explained either by constitutional characteristics, disposition characteristics, qualification and competence characteristics or by adaptation characteristics of a worker [19]. Constitutional characteristics (e.g. gender, body type, nationality etc.) are invariable in a person's life-cycle so that it is difficult to change a person's ability to cope with the required task complexity. This also applies to disposition characteristics (e.g. personality, age, intelligence, body weight etc.) which can be changed over time – even though not directly by the employee. However, qualification and competence characteristics (e.g. experience, knowledge, skills, education, competence etc.) can be varied through long, medium and short-term work structuring processes. For instance, dedicated training can help to ensure that a work task is no longer perceived as too complex, i.e. more complex than the complexity optimum of the parabolic curve, by the individual worker. In this context, Dalhöfer and Prieß [15] regard the line manager as key figure in complexity management, as the complexity optimum could individually be designed for every employee. Finally, short-term work structuring measures may lead to a variation of adaptation characteristics (reaction to energetic-effector, informational-mental or emotional stress; e.g. strain, fatigue, satisfaction etc.) and to a reduction of subjectively perceived task complexity, e.g. through supporting information systems.

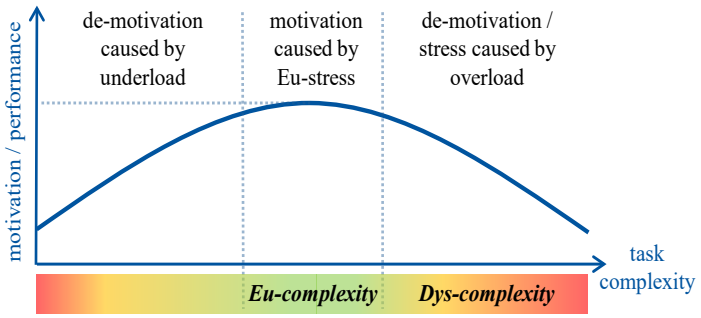


Figure 14: Performance, motivation and task complexity according to Dalhöfer and Prieß [15]

This article makes the first attempt to extend these conceptual considerations towards cooperative forms of work and hence further developing the concept of Dalhöfer and Prieß [15]. For the example of a work group, it is assumed that the original concept for the individual worker still is valid. Thus, each work group has individual complexity parabolas for each person in this group. The individual complexity parabolas are now transferred into a three-dimensional diagram, which contains the individual persons of the work group as a third axis. This creates a three-dimensional graph that can be covered with an envelope curve. In this article, the concept of a "task complexity mountain range" of a work group is introduced as a metaphor. This concept is illustrated in Figure 15.

It is evident that although the task complexity mountain range is composed of the individual graphs, the interval of a complexity optimum also exists for the work group as a whole. The interval appears through the envelope curve, which covers the individual graphs. However, this optimum cannot clearly be determined due to the different apex points of the individual parabolas and can only be described as an interval in which the abilities of the group members complement each other. Within this interval, the overall complexity of the group's work task needs to be distributed adequately according to the individual complexity parabolas of the particular group members. However, identifying the optimum is already a complex task, as the optimum may shift dynamically depending on the group structure and internal group processes.

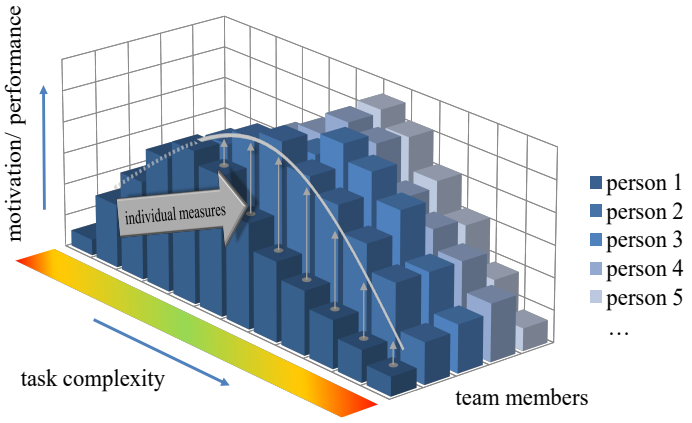


Figure 15: Development of the individual task complexity concept of Dalhöfer and Prieß [15] towards a "task complexity mountain range" for cooperative forms of work; including individual coaching measures to adjust the complexity curve of person 1

In the context of cooperative forms of work, the concept of leadership in complexity management of Dalhöfer and Prieß [15] is even more important: within the group, the line manager has to respond to individual working persons and individually adjust the complexity optimum. In Figure 15, the arrow labelled "individual measures" indicates that through dedicated training the complexity optimum of a particular person – and in a reverse conclusion of the entire work group – can be shifted towards more complex work tasks. For instance, this is achievable by systematically enhancing the individual qualification and competence characteristics. In addition, the work group as a whole can be enabled by the line manager to handle more complex tasks independently and to react adaptively to unpredictable situations. Transformational leadership, for example, can achieve this by strengthening the intrinsic motivation of employees through transforming their values and ambitions as well as supporting the individual development [67]. ElMaraghy et al. [17] point out that collaboration is able to lead to adaptive and creative forms of work that can handle complexity very effectively.

However, the expert interviews have shown that the transformation towards delegating and cooperative management initially increases complexity – and that sufficient complexity handling can only be achieved in the target state of the transformation process. Also, participation was identified in the expert interviews as most frequently mentioned measure for managing complexity in digitalized work systems. This should consequently be a subject of research in the context of management of work groups. With regard to management processes, goal-setting theory [68] should also be examined in greater depth. For instance, the effect of conflicting goals on work processes of groups with a high degree of autonomy should be inquired. The expert interviews have indicated that a higher degree of autonomy and conflicting objectives increase complexity for working persons and that these aspects are intensified in the course of digitalization. Therefore, further research is needed to determine how a good target quality can be achieved for work groups. Overall, especially in the context of digitalization, an intensive investigation of cooperative leadership processes in group work is particularly relevant in order to develop strategies for complexity management for work groups.

According to Schuh and Riesener [65], corporate complexity management strategies for optimizing the product portfolio not only include measures to handle complexity, but also measures to avoid and reduce complexity. These could also be applied to complexity management of group tasks. Thus, there are two further strategies for designing the "optimal" task complexity for work groups: on the one hand, task complexity can be reduced via redesigning the work process (corrective approach), if the level of task complexity lies within the interval that decreases work performance and motivation due to overstraining conditions. On the other hand, task complexity is initially avoidable in terms of preventive work design.

The degree of optimal task complexity of a work group or team also always depends on the operational conditions. As an example, resources provided, support from other organizational units or digital support tools determine how well complex tasks can be solved by the work group. However, the use of digital support systems also requires a differentiated consideration: the expert interviews have shown that digital assistance systems are able to reduce complexity if they are designed in a user-centered manner. In addition to participatory development processes, which have the potential to increase acceptance of new systems, there are instructions for ergonomic designs of such systems in norms, e.g. in EN ISO 9241 [69]. Nevertheless, the interview results also showed that perceived complexity of a work task may increase due to user-unfriendly designs of digital work tools. In addition, the perceived complexity depends on the life-cycle phase of the support system. It is important to distinguish whether the system is currently being introduced in the company, adapted or already established. Even though a new digital support system may be intended to make complexity manageable in a work system, the introduction of a new system will initially generate a (subjective) increase in complexity. Initially, the system is unfamiliar and requires training. This phase is followed by an iterative phase in which the employees become familiar with the use of the system. At the same time, possible system errors are eliminated and the system is adjusted to the needs of the employees. In the third phase, the "steady state", the intended subjective complexity reduction of the work task is sustainably realized. The interviews indicated that a participatory development and testing procedure of such systems facilitates to reach this target state in a shorter period of time. To accompany this, appropriate training for the use of the systems should support the introduction process. Figure 16 illustrates the settling process of digital support systems with regard to perceived task complexity.

Overall, it is apparent that considerations relating to complexity and digitalization in the context of cooperative forms of work offer multifaceted research questions. In order to contribute to bridging the identified research gaps, a holistic empirical analysis of the interrelationships in a structural equation model is suggested (see e.g. [59]). Further research steps and a general discussion are considered in the following section.

6. General discussion and future research steps

Diverse disciplines conduct research on the construct of complexity with specific points of interest and with different methods. This article explored complexity drivers that arise through digitalization. The qualitative content analysis of the

interviews identified various complexity drivers in digitalized labor systems in all six organizational dimensions. They may be employed to enhance current complexity driver systematizations that already exist in literature.

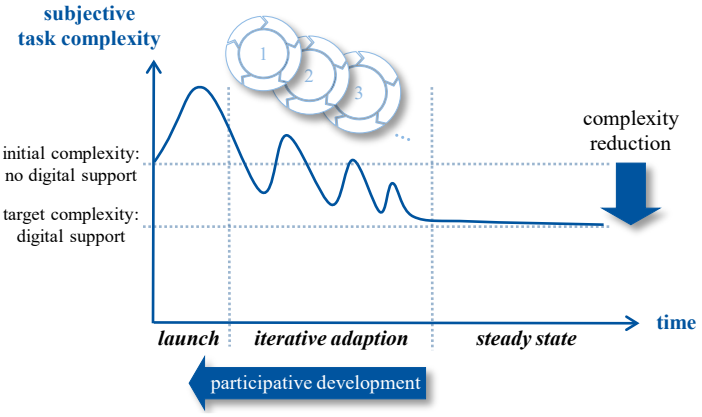


Figure 16: Settling process of digital support systems with respect to subjective task complexity

Since digitalization changes labor systems and whole industrial sectors, it is crucial to gain a deep understanding of which factors raise the level of complexity not only from a technical but at the same time from a work design perspective. The findings show that the supply of lots of information as well as diverse options of displaying them may imply complexity for employees in a negative connotation. This is why it is crucial to agree on guidelines for company-internal communication. A lower level of complexity is obtainable by reducing the number of communication systems. Moreover, information should be organized and supplied in a manner that makes is easy for operators to spot the information that is needed by them to perform their jobs.

However, it is frequently based on the viewpoint whether a certain topic should be seen as a driver of complexity. In this connection, the complexity drivers belong to various points of view which relate to whole systems as well as to subjectively perceived complex incidences from the perspective of an employee. Regarding the operator's perspective, a new digital support system could reduce complexity for an operator, if it was well designed. Yet, complexity rises on a system level, as the amount of system elements and system maintenance procedures increases.

Furthermore, it should be mentioned that the organizational dimensions cannot be examined separately, since they are interdependent in diverse ways. This is why it is especially complicated to assign drivers to particular dimensions. The validity of the results of the analysis procedure can be increased by enlarging the sample size of interviewed experts. This could enable cross-references with the results of the interviews as well as complexity drivers could be identified that are specific for certain industrial sectors. All in all, the conducted method of semi-structured interviews with experts enabled to qualitatively analyze the answers of the interviewees. Therefore, this method was adequate for the first research step of exploration. If conducted with sufficient many persons for the coding procedure, the method may offer reliable results. However, the inter-code-reliability of the results can be enhanced and quantitatively judged by repeating the analysis procedure with more than two coders. Further research steps will use the findings to develop theories and to conduct ongoing quantitative empirical studies.

This article discussed the findings in the context of cooperative forms of work and offers initial concepts that will be examined with empirical research methods in future. The studies will have to show whether the concept of the task complexity mountain range should be retained or adapted. Furthermore, the effects of digitalization on output and process measures of cooperative forms will be assessed.

Eventually, complexity should not be seen unequivocally as a thing that should be avoided at all costs. Actually, complexity is also the basis for market leadership, if enterprises have an expertise in reproducing the demanded market complexity. However, the important question is how much capacity is required to fulfill these requirements so that too complex organizations can be prevented. This especially leads to the overall challenge of how work systems should be designed to provide rather challenging than overstraining work conditions for employees.

7. Conclusion

This article examined which complexity drivers in labor systems are caused by digitalization. Initially, diverse complexity definitions from relevant research disciplines were presented. General properties of complexity in theory were identified. Moreover, the used method of semi-structured interviews with experts for exploring the complexity drivers was outlined. The group of interviewees was made up of 23 experts with positions in German economy and science. The findings show that the digital transformation could remarkably alter work processes and whole industrial sectors. In this connection, diverse complexity drivers in digitalized labor environments were explored and clustered into six organizational dimensions. The findings can be employed to point out far-reaching consequences of digitalization, because the six organizational dimensions and the topics of complexity and digitalization are mutually interrelated. Further, measures for handling complexity in digitalized work systems that were extracted from the interviews. The findings were discussed in the context of cooperative forms of work as well as existing concepts for explaining complexity impacts were transferred to this topic. In this context, the concept of the task complexity mountain range for cooperative forms of work was developed. Future research steps will attempt to empirically assess the complexity in digitalized labor systems. In this context, it will be analyzed which effects complexity and digitalization have on cooperative forms of work in production systems. The results may be used for developing additional measures to handle complexity in digitalized work systems to those that were presented in this article.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research and development project "TransWork" is funded by the German Federal Ministry of Education and Research (BMBF)

within the program "Innovationen für die Produktion, Dienstleistung und Arbeit von morgen" according to Grant No. 02L15A162 supervised by the Project Management Agency of Karlsruher Institute of Technology (PTKA). The authors are responsible for the content of this publication. The authors would like to express their gratitude for this support.

Appendix

Appendix 1: Questionnaire guide for the semi-structured interviews

1. Name, company, company size, function and tasks in company

2.1 How has digitalization changed your company and business in the last 10 years?

2.2 How will digitalization change your company and business in the next 10 years?

3. How does digitalization influence the level of complexity in the 6 dimensions of work organization? Which aspects of digitalization cause a higher level of complexity?

a) Process organization (e.g. production planning, material flow, information flow, innovative assembly concepts)

b) Organizational structure (e.g. hierarchy levels, delegation of responsibility), decision-making powers)

c) Technology (e.g. production technology, information technology)

d) Working conditions / incentive systems (e.g. employment contracts, work safety, working hours, remuneration)

e) Product (e.g. product design, product structure, new business models)

f) Personnel (e.g. cooperation, communication, leadership, personnel planning, qualification)

4. What measures can be developed to manage the complexity caused by digitalization?

References

[1] B. A. Latos, M. Harlacher, P. M. Przybysz, S. Mütze-Niewöhner, "Transformation of Working Environments Through Digitalization: Exploration and Systematization of Complexity Drivers" in IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, Singapore, 2017. http://dx.doi.org/10.1109/IEEM.2017.8290059

[2] G. Schuh, M. Riesener, J. Kantelberg, N. Steireif, "Transmission of software-related agile mechanisms of action towards product development processes for technical products" in IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, Singapore, 2017. https://doi.org/10.1109/IEEM.2017.8290192

[3] C. Châlons, "Die Rolle der IT als Enabler für Digitalisierung", In: F. Abolhassan (Ed.), Was treibt die Digitalisierung? Warum an der Cloud kein Weg vorbeiführt, Wiesbaden, Germany: Springer, 27-37, 2016. http://dx.doi.org/10.1007/978-3-658-10640-9

[4] C. Aichele, M.. Schöneberger, Die Digitalisierung der Energiewirtschaft: Potenziale und Herausforderungen der IKT-Branche für Utility 4.0, In: O. D. Doleski (Ed.), Herausforderung Utility 4.0. Wie sich die Energiewirtschaft im Zeitalter der Digitalisierung verändert, Wiesbaden, Germany: Springer, 501-530, 2017. https://doi.org/10.1007/978-3-658-15737-1

[5] W. Widuckel, K. de Molina, M. J. Ringlstetter, D. Frey, Arbeitskultur 2020. Herausforderungen und Best Practices der Arbeitswelt der Zukunft, Wiesbaden, Germany: Springer, 2015. https://doi.org/10.1007/978-3-658-06092-3

[6] W. M. Walter, "Der Einfluss der Digitalisierung auf die Organisation eines Unternehmens", In: O. D. Doleski (Ed.), Herausforderung Utility 4.0. Wie sich die Energiewirtschaft im Zeitalter der Digitalisierung verändert, Wiesbaden, Germany: Springer, 227-248, 2017. https://doi.org/10.1007/978-3-658-15737-1

[7] W. Bauer, S. Schlund, "Wandel der Arbeit in indirekten Bereichen – Planung und Engineering", In: H. Hirsch-Kreinsen, P. Ittermann, J. Niehaus (Ed.), Digitalisierung industrieller Arbeit. Die Vision Industrie 4.0 und ihre sozialen Herausforderungen, 1st ed., Berlin, Germany: Nomos, 53-70, 2015. https://doi.org/10.5771/9783845263205-1

[8] Institut DGB-Index Gute Arbeit, DGB-Index Gute Arbeit. Der Report 2016. Wie die Beschäftigten die Arbeitsbedingungen in Deutschland beurteilen, Berlin, Germany: PrintNetwork pn / ASTOV Vertriebsgesellschaft mbH, 2016.

[9] O. Stettes, Arbeitswelt der Zukunft. Wie die Digitalisierung den Arbeitsmarkt verändert, Cologne, Germany: IW Medien, 2016.

[10] H. Hirsch-Kreinsen, "Digitization of industrial work. Development paths and prospects" Journal for Labour Market Research, 49(1), 1-14, 2016. https://doi.org/10.1007/s12651-016-0200-6

[11] L. Bertschek, B. Dworschak, P. Meil, T. Niebel, J. Ohnemus, T. Vetter, H. Zaiser, Arbeitsmarkt 2030 - Digitalisierung der Arbeitswelt. Fachexpertisen zur Prognose 2016, Munich, Germany: Economix Research & Consulting, 2016. https://doi.org/10.3278/6004559w

[12] S. Schlund, B. Pokorni, Industrie 4.0 - Wo steht die Revolution der Arbeitsgestaltung? Ergebnisse einer Befragung von Produktions-verantwortlichen deutscher Unternehmen, Ulm, Germany: Ingenics AG, 2016.

[13] B. A. Latos, M. Harlacher, M. El-Mahgary, D. Götzelmann, P. M. Przybysz, S. Mütze-Niewöhner, C. Schlick, "Komplexität in Arbeitssystemen: Analyse und Ordnung von Beschreibungsansätzen aus unterschiedlichen Disziplinen" in 63th Frühjahrskongress der Gesellschaft für Arbeitswissenschaft FHNW, Brugg-Windisch, Switzerland, 2017. https://doi.org/10.18154/RWTH-2017-04217

[14] M. Schwemmle, P. Wedde, Digitale Arbeit in Deutschland: Potenziale und Problemlagen, Bonn; Germany: bub Bonner Universitäts-Buchdruckerei, 2012

[15] J. Dalhöfer, M. Prieß, "Führung im Komplexitätsmanagement: Zusammenhang zwischen Komplexität, Motivation und Stress" ZWF Zeitschrift für Wirtschaftlichen Fabrikbetrieb, 107(1-2), 87-93, 2012. https://doi.org/10.3139/104.110666

[16] A. Hoeschen, "Complexity-based distribution of value added in multinational production systems" (Thesis in German) Ph.D. dissertation, RWTH Aachen University, Germany, 2015.

[17] W. ElMaraghy, H. ElMaraghy, T. Tomiyama, L. Monostori, "Complexity in engineering design and manufacturing" CIRP Annals – Manufacturing Technology 61(2), 793-814, 2012. https://doi.org/10.1016/j.cirp.2012.05.001

[18] R. Riedl, Strukturen der Komplexität: Eine Morphologie des Erkennens und Erklärens, Berlin, Germany: Springer, 2000. https://doi.org/10.1007/978-3-642-56946-3

[19] C. Schlick, R. Bruder, H. Luczak, Arbeitswissenschaft, 4th ed., Heidelberg, Germany: Springer, 2018. https://doi.org/10.1007/978-3-662-56037-2

[20] C. R. Shalizi, "Methods and Techniques of Complex Systems Science", In: T. S. Deisboeck, J. Y. Kresh (Ed.) Complex systems science in biomedicine, New York, USA: Springer, 49-131, 2006.

[21] H. A. Simon, The sciences of the artificial, 3rd ed., Cambridge, USA: MIT Press, 1996.

[22] M.-O. Blockus, "Komplexität in Dienstleistungsunternehmen: Komplexitätsformen, Kosten- und Nutzenwirkungen, empirische Befunde und Managementimplikationen", Ph.D. dissertation, Universität Basel, Switzerland, 2010, 1st ed., Wiesbaden, Germany: Gabler Research, 2010. https://doi.org/10.1007/978-3-8349-8958-1

[23] L. von Bertalanffy, "Vorläufer und Begründer der Systemtheorie" in System Theory Research and Information, (Systemtheorie Forschung und Information), R. Kurzrock, Berlin, Germany: Colloquium-Verlag, 17-28, 1972.

[24] J. B. Rosser, "Computational and Dynamic Complexity in Economics", In: J. B. Rosser (Ed.), Handbook of Research on Complexity, Cheltenham, UK: Edward Elgar Publishing, 2009. https://doi.org/10.4337/9781781952665

[25] G. Schuh, U. Schwenk, Produktkomplexität managen. Strategien, Methoden, Tools, München, Germany: Carl Hanser, 2001.

[26] H. Ulrich, G. J. Probst, Anleitung zum ganzheitlichen Denken und Handeln. Ein Brevier für Führungskräfte, Bern, Switzerland: Haupt, 1988.

[27] A. V. Deshmukh, J. J. Talavage, M. M. Barash, "Complexity in manufacturing systems, Part 1: Analysis of static complexity" IIE Transactions, 30(7), 645-655, 1998.

[28] O. Kuzgunkaya, H. A. ElMaraghy, "Assessing the structural complexity of manufacturing systems configurations" International Journal of Flexible Manufacturing Systems, 18(3), 145-171, 2006.

[29] S. Sivadasan, J. Efstathiou, A. Calinescu, L. H. Huatuco, "Advances on measuring the operational complexity of supplier–customer systems"

European Journal of Operational Research, 171(1), 208-226, 2006. https://doi.org/10.1016/j.ejor.2004.08.032
[30] S. Mattsson, P. Gullander, U. Harlin, G. Bäckstrand, Å. Fasth, A. Davidsson, "Testing Complexity Index – a Method for Measuring Perceived Production Complexity" in 45th CIRP Conference on Manufacturing Systems, Athens, Greece, 394-399, 2012. https://doi.org/10.1016/j.procir.2012.07.068
[31] L. Zeltzer, V. Limère, H. van Landeghem, E.-H. Aghezzaf, J. Stahre, "Measuring complexity in mixed-model assembly workstations" International Journal of Production Research, 51(15), 4630-4643, 2013. https://doi.org/10.1080/00207543.2013.783246
[32] D. Dörner, H. W. Kreuzig, F. Reither, T. Stäudel, Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität, Bern, Switzerland: Hans Huber Verlag, 1983.
[33] D. Dörner, Die Logik des Misslingens, Strategisches Denken in komplexen Situationen, 10th ed., Hamburg, Germany: Rowohlt Verlag, 2011.
[34] A. Kluge, Wissenserwerb für das Steuern komplexer Systeme, Lengerich, Germany: Wolfgang Pabst Science Publishers, 2004.
[35] H. Willke, Systemtheorie 1: Grundlagen, 5th ed., Stuttgart, Germany: UTB, 1996.
[36] N. Luhmann, Soziale Systeme. Grundriß einer allgemeinen Theorie, Frankfurt a. M., Germany: Suhrkamp Verlag, 1987.
[37] W. Rohmert, "Das Belastungs-Beanspruchungs-Konzept" Zeitschrift für Arbeitswissenschaft, 38(4), 193-200, 1984.
[38] H. Luczak, "Untersuchungen informatorischer Belastung und Beanspruchung des Menschen" Fortschrittsberichte der VDI-Zeitschriften, 10(2), Düsseldorf, Germany: VDI-Verlag, 1975.
[39] M. W. McElroy, "Integrating complexity theory, knowledge management and organizational learning" Journal of Knowledge Management, 4 (3), 195-203, 2000. https://doi.org/10.1108/13673270010377652
[40] W. Vogel, R. Lasch, "Complexity drivers in manufacturing companies: A literature review" Logistics Research, 9(1), 1-66, 2016. https://doi.org/10.1007/s12159-016-0152-9, licensed under: http://creativecommons.org/licenses/by/4.0/
[41] K.-P. Schoeneberg, Komplexitätsmanagement in Unternehmen. Herausforderungen im Umgang mit Dynamik, Unsicherheit und Komplexität meistern, Wiesbaden, Germany: Springer Gabler, 2014. https://doi.org/10.1007/978-3-658-01284-7
[42] S. Bednar, V. Modrak, "Mass Customization and its Impact on Assembly Process' Complexity" International Journal for Quality Research, 8(3), 417-430, 2014
[43] N. Asadi M. Jackson, A. Fundin, "Drivers of Complexity in a Flexible Assembly System - A Case Study" Procedia CIRP, 41, 189-194, 2016. https://doi.org/10.1016/j.procir.2015.12.082
[44] N. Asadi, S. Javadi, "Key complexity dimensions in assembly systems with mixed-model assembly lines - a multiple case study" in 5th World Conference on Production and Operations Management, Havana, Kuba, 2016.
[45] S. Mattsson, M. Karlsson, P. Gullander, H. van Landeghem, L. Zeltzer, V. Limère, E.-H. Aghezzaf, "Comparing quantifiable methods to measure complexity in assembly" International Journal of Manufacturing Research, 9(1), 112-130, 2014. https://doi.org/10.1504/IJMR.2014.059602
[46] P. Mayring, Einführung in die qualitative Sozialforschung, 5th ed., Weinheim, Germany: Beltz, 2002.
[47] M. Pfadenhauer, Professionelles Handeln, Wiesbaden, Germany: VS Verlag für Sozialwissenschaften, 2005.
[48] P. Mayring, Qualitative Inhaltsanalyse: Grundlagen und Techniken, Weinheim, Germany: Beltz, 2010.
[49] P. Mayring, Qualitative content analysis: theoretical foundation, basic procedures and software solution. Klagenfurt, Germany: Beltz, 2014.
[50] J. Gläser, G. Laudel, Experteninterviews und qualitative Inhaltsanalyse, 3rd ed., Wiesbaden, Germany: VS Verlag für Sozialwissenschaften, 2009.
[51] A. Petz, "Produktivitätsbewertung und simulationsbasierte Gestaltung von wissensintensiven Dienstleistungen" Ph.D. dissertation, RWTH Aachen University, Germany, 2017. https://doi.org/10.2370/9783844057058
[52] A. Petz, S. Duckwitz, C. Schmalz, "Productivity of Services: An Explorative Study in the Electrical and Chemical Engineering Sector" Amfiteatru Economic, 14(Special No. 6), 635-652, 2012.
[53] S. Duckwitz, A. Petz, C. Schmalz, R. Jahnel, P. Lehmacher, C. M. Schlick, "Dienstleistungsproduktivität aus Unternehmensperspektive - Ergebnisse semistrukturierter Leitfadeninterviews", In: C. M. Schlick, M. Schenk, D. Spath, W. Ganz (Ed.), Produktivitätsmanagement von Dienstleistungen: Modelle, Methoden und Werkzeuge, Berlin, Germany: Springer Vieweg, 13-27, 2016. https://doi.org/10.1007/978-3-642-45071-6
[54] J. Wegge, F. Jungmann, S. C. Liebermann, K.-H. Schmidt, B. C. Ries, "Altersgemischte Teamarbeit kann erfolgreich sein: Empfehlungen für eine ausgewogene betriebliche Altersstruktur" Sozialrecht + Praxis, 7, 433-442, 2011.
[55] M. Högl, "Teamarbeit in innovativen Projekten: Einflussgrößen und Wirkungen" Ph.D. dissertation, Universität Karlsruhe, Germany, 1998.
[56] M. Högl, H. G. Gemünden, "Teamwork Quality and the Success of Innovative Projects: A Theoretical Concept and Empirical Evidence" Organization Science, 12(4), 435-449, 2001. https://doi.org/10.1287/orsc.12.4.435.10635
[57] T.-M. Myskovszky von Myrow, "Teamwork during innovation projects: The effects of diversity, participation, and the organizational context" Ph.D. dissertation, RWTH Aachen University, Germany, 2018.
[58] P. M. Przybysz, "Empirische Untersuchungen der Auswirkung von Altersdiversität in Innovationsteams" Ph.D. dissertation, Ernst-Moritz-Arndt-Universität Greifswald, Germany, 2016.
[59] R. A. Stranzenbach, "Analysis and Design of Group Work in Integrated Production Systems" (text in German) Ph.D. dissertation, RWTH Aachen University, Germany, 2018. in press
[60] S. Tannenbaum, R. Beard, E. Salas, "Team building and its influence on team effectiveness. An examination of conceptual and empirical developments" Advances in Psychology, 82, 117-153, 1992. https://doi.org/10.1016/S0166-4115(08)62601-1
[61] P. Trompisch, "The implications of Industry 4.0 on the future of work" (article in German with an abstract in English) Elektrotechnik & Informationstechnik 134(7), 370-373, 2017. https://doi.org/10.1007/s00502-017-0531-1
[62] J. Rasmussen, A. M. Pejtersen, L. P. Goodstein, Cognitive Systems Engineering, New York, USA: Wiley, 1994.
[63] S. Mattsson, "What is perceived as complex in final assembly? To define, measure and manage production complexity" Ph.D. Thesis, Chalmers University of Technology Gothenburg, Sweden, 2013.
[64] T. Fässberg, U. Harlin, K. Garmer, P. Gullander, Å. Fasth, S. Mattsson, K. Dencker, A. Davidsson, J. Stahre, "An Empirical Study Towards a Definition of Production Complexity" in 21st International Conference on Production Research, Stuttgart, Germany, 2011.
[65] G. Schuh, M. Riesener, Produktkomplexität managen. Strategien –Methoden – Tools, 3rd ed., Munich, Germany: Hanser, 2017. https://doi.org/10.3139/9783446453340
[66] A. Kaiser, "Integriertes Variantenmanagement mit Hilfe der Prozesskostenrechnung" Ph.D. Thesis, University of St. Gallen, Switzerland, 1995.
[67] S. D. Dionne, F. J. Yammarino, L. E. Atwater, W. D. Spangler, "Transformational leadership and team performance" Journal of Organizational Change Management, 17(2), 177-193, 2004. https://doi.org/10.1108/09534810410530601
[68] E. A. Locke, G. P. Latham, A Theory of Goal-Setting and Task Performance, New Jersey, USA: Englewood Cliffs, Prentice Hall, 1990.
[69] EN ISO 9241 Part 110 (September 2008): Ergonomics of human-system interaction - Part 110: Dialogue principles

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 173-186 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Supporting Active Aging Through A Home Automation Infrastructure for Social Internet of Things

Vittorio Miori*, Dario Russo, Luca Ferrucci

Institute of Information Science and Technologies (ISTI), CNR-National Research Council of Italy, Via Moruzzi 1, Pisa 56124, Italy

ARTICLE INFO

Article history:
Received: 30 May, 2018
Accepted: 05 July, 2018
Online: 29 July, 2018

Keywords:
SIoT
IoT
Interoperability
SOA
SOCIALIZE
AAL
Smart Home
Active Ageing

ABSTRACT

The number of smart devices inside the living environments grows more and more every day, and with it the need to interface them in order to share data and activate functions. Each day, new scenarios and new applications emerge to make our lives easier in many different contexts. The AAL SOCIALIZE project aims to create a new reference platform for active ageing by applying technological solutions to care for the health of the elderly, facilitate their daily activities, simplify their access to dedicated services.
The work presented herein describes the "Elderly Monitoring service" – an IoT module of the SOCIALIZE platform, whose aim is to monitor elderly people's activities and physiological parameters and to combine smart device functions. This modeling process can actively contribute to anticipating, and thereby preventing, emergency situations, increasing elderly autonomy and making them feel "protected" and "safe".
In this regard, the system implements a practicable, scalable solution that allows heterogeneous devices belonging to different domotic systems using diverse protocols to interact with each other, share data and services and bind together through 'friendship', thereby creating an infrastructure suitable for the creation of a true Social IoT (SIoT).

1. Introduction

This paper is an extension of the work originally presented in *2017 Global Internet of Things Summit* (*GIoTS*) [1].

The issues involved in welfare and health costs sustainability are currently the focus of widespread discussion in many nations of the world. Of such issues, one of the most frequently addressed is population ageing. The ageing population will soon become one of the most demanding Big Societal Challenges that the world will face [2]. Over the last 150 years, life expectancy has risen by 50 years, and over the last half century alone it has increased by three years every decade.

Hand in hand with ageing, it is highly likely that the elderly will experience an increase in disability and dependence, along with a greater incidence of age-related illnesses and the need for medication [3]. The majority of the ageing population experiences progressive deterioration of their health from 50 to 80 years of age. As a result, there is an interval of 30 years during which subjects age with deteriorating health – a critical aspect to take into consideration when seeking to preserve good quality of life for the elderly. Over recent years the elderly population has been growing. This is mainly due to the progressive increase in life expectancy and the low fertility rates affecting an ever greater proportion of the overall population [4]. This progressive increase in the elderly population has led to a new phenomenon: so-called active ageing. According to the definition of the *World Health Organization* [5], active aging is "a process of optimization of opportunities for health, participation and safety in order to improve the quality of life of older people". Active aging is epitomized by elderly people who, rather than spend their entire day sitting at home, bored, in an armchair, voluntarily stay in the workforce for as long as they can, they exercise, they train, they take care of their children and grandchildren. They are over 65-70 years of age, they are healthy, they are satisfied and they often have acquired skills, including on the technological and logical levels. In the near future, the upcoming candidates to join the ranks of the active aging are currently in the age range of 50-60 years. The European Community is actively involved in caring for the active ageing, and to such end has launched the programme "Ageing well in the information society". Moreover, the European Commission declared 2012" The Year of Active Ageing and Solidarity between Generations". One result of the measures adopted in 2012 is the formulation of the active ageing index, which measures the ability of older people to achieve their full potential in terms of

* Corresponding author: Vittorio Miori, Email: vittorio.miori@isti.cnr.it

employment, social and cultural participation and maintaining autonomy.

In an information society, *Information and Communication Technology* (*ICT*) plays an important role for the active elderly. Utilizing the *WWW*, for example, they can get information, have social interactions, participate in social and political life, enjoy simplified access to administrative and commercial transactions, find online entertainment and leisure activities, and so on [6]. Moreover, *ICT* offers a great opportunity for older people living home alone. According to the *Internet of Things* (*IoT*) paradigm, objects are rendered 'intelligent'– they can be discovered, localized, and acquire, process and exchange data. Smart Home & Building applications are particularly important in the *IoT* scenario, as they are the link between the individual (citizen, consumer) and the overlying layers implementing the *IoT* paradigm (Smart City, Smart Grid) [7].

The *IoT* is however not merely a distant vision of the future – it is already here and is having an impact on more than just technological developments. It permits sophisticated devices to share information directly with each other and the cloud, making it possible to collect, record and analyze new data streams faster and more accurately. So, it is now possible to exploit a broad set of services able to improve the autonomy of the elderly [8] and constantly monitor their health status and identify any potentially hazardous conditions [9]. Connected 'smart objects' can be used to collect data on the health status of home residents, particularly those at risk. The *IoT* will thus make possible to create environments equipped with such intelligent devices in which an older person's state of health is constantly monitored, allowing for more effective management of his or her long-term well-being. The advent of the *IoT* and its new technological capabilities offers many advantages over traditional health monitoring. In particular, conventional monitoring systems are not predictive, that is, they cannot offer the same accuracy as modern biometric detection systems, which are capable of recognizing emergencies before they occur, and they do not have the ability to "automatically" launch a warning to formal and/or informal caregivers without direct action by the elderly person.

Instead, by combining wearable devices with *IoT* these limitations can be overcome. Commercial products offering similar such features already exist: intelligent watches whose functions range from detecting movements indicating a possible fall and contacting the authorities or relatives, to the ability to "learn" a person's normal activities and send an alert by text message, email, or other app if a significant change occurs (e.g. a user lies in bed for an entire weekend).

The last frontier of the *Internet of Things* is the evolution of objects that are 'merely' smart to objects that can be social. This paradigm is called *Social Internet of Things* (*SIoT*) [10,11]. The *SIoT* is defined as an *IoT* in which objects are capable of establishing social relationships with other objects, independently of human intervention, thereby creating a social network of objects.

The key principle guiding the *SOCIALIZE* project [12] is to create a complete technological solution which effectively fulfills the general objectives of the third call of the *EU – AAL* (Ambient Assisted Living) programme. The *SOCIALIZE* project is creating a new, specialized platform expressly dedicated to today and tomorrow's active ageing people in order to facilitate their participation in social life.

The authors' main contribution to the project presented in this work is the study and development of a *SIoT* ready service, called the *Elderly Monitor*, which is fully integrated into the *SOCIALIZE* platform. The function of the service is to monitor the activities of home occupants, collect and analyze personal and environmental parameters from various, technologically heterogeneous devices and enable them to cooperate and share data. The results of the monitoring can be accessed by doctors, relatives and/or caregivers as deemed appropriate with regard, for example, to the cognitive, health status and wishes of the person in question.

The paper is organized as follows: in *The Social IoT (SIoT)* section is introduced the feature and advantages of the *SIoT* apprach, in the *Related work* section are shown the most significant solutions for the monitoring and analysis of the user's health parameters, the main applications of the *IoT* in the health field and the most widely used *SIoT* approaches. In the section *Socialize Project* is presented the *SOCIALIZE* project with its main features and objectives. This section allows to illustrate the context in which the application presented in this work was studied and developed. In the section *SmartSMILE framework* is described in detail the ideas of the proposed solution. In the *Implementation* section are described how some domotic technologies have been integrated in the framework. In the *Field trials* section are shown the deployment and the experiments carried out to verify the correct functioning of the system. Finally, we have the *Conclusions* section that summarizes and concludes the work.

2. The Social IoT (SIoT)

Adoption of the *SIoT* paradigm presents several further advantages: (*i*) the resulting structure of such object networks can be shaped as required to guarantee network navigability and effectively perform the discovery of objects and services, thereby guaranteeing scalability, just as in human social networks; (*ii*) a level of trustworthiness can be established to leverage the degree of interaction among things that are 'friends'; (*iii*) models designed to study social networks can be reused to address *IoT* related issues (intrinsically related to extensive networks of interconnected objects). Thanks to the implementation of this paradigm, devices become part of a social network of objects. As in human social networks, the concept of "friend" plays a key role in *SIoT*. By exploiting device friendship, things can be made to interoperate autonomously by implementing intelligent behaviors [13]. This scenario has opened up the possibility for new object intelligence: the ability to discover new services, activate new contacts with other objects, share information with them, and use the capabilities of other objects towards the attainment of a common goal.

In order for *SIoT* principles to be effectively applied, the communication and integration infrastructure of *IoT* devices must be built in such a way as to address specific challenges. If we consider the domestic environment, most of them concern the lack of interoperability between the diverse installed home automation systems, but not only. The *SIoT* must be based on an infrastructure able to support the following services:

- *Discovery*: the ability for devices to detect the functions of other devices. This service is useful to automatically identify device functions that could be combined together to create new, more advanced ones. This implies that each service of each device must be identified, addressed and described using an ontology, taxonomy or free-text description that follows predefined rules in order to be computed.

- *Composition*: the ability to combine the functions of different devices to create new system capabilities. To do this, the constituent devices of a system have to "speak" the same language. This can only be obtained either by using devices that adopt the same communication technology or by implementing interoperability between natively incompatible devices to share states, values and function calls. Interoperability is the ability of two or more incompatible systems to communicate in a way to create a functional dependence between them, allowing them to act like a single system that behaves as a single entity [14].
- *Trustworthiness management*: the ability to understand how to use the information provided by other devices. To implement this service, it is useful to apply the concepts of centrality, prestige and reliability. Centrality measures the number of contacts any given device has within the network; prestige measures how many other devices seek that device. Centrality and prestige are summarized in a value called reliability. These measures are particularly interesting in big *IoT* scenarios, in which the network is made up of numerous devices and a mechanism is needed to determine which are most dependable for reaching a particular goal.

On the contrary, the vast majority of existing *IoT* interoperability platforms are not designed to offer all these features in an integrated way. Devices and applications do not communicate directly between each other, but use cloud connectivity, computing and storage features as intermediaries.

In fact, in January 2016, with the *Horizon 2020 research programme*, the European Commission funded seven projects to address the interoperability of the *Internet of Things*: *Inter-IoT* [15], *BIG IoT* [16], *AGILE* [17], *symbIoTe* [18], *VICINITY* [19], *bIoTope* [20]. In general, the most widely followed technical approach to address these issues is the implementation of a hierarchical *IoT* stack connecting smart objects and *IoT* gateways with the cloud. In these projects, interoperability is implemented by collecting device data in the cloud (typically in a database) where they can be queried in order to process information.

3. Related work

The current scientific literature addresses the issue of elderly health monitoring with solutions that are able to collect and then display data, typically through a web interface, to formal and / or informal caregivers who can check them to oversee an elderly person's health status. In addition to allowing viewing of the collected data, most of the proposed solutions include a preliminary automatic analysis of the occupant's health. Such analyses take different approaches. Some check whether a prefixed threshold for some specific health parameter (e.g. heart rate) has been exceeded and then send a notification to designated recipients. Others use more advanced intelligent systems that are able to understand trends in a set of specific user parameters and thereby make decisions and automatically launch alarms and warnings when vital sign values are considered suspicious. Some examples of this kind of systems are those that analyze data in real time. In [21], the author proposes a pervasive patient health monitoring (*PPHM*) system based on wireless wearable devices integrated through a cloud computing infrastructure. The work presents also a solution based on *ECG* to monitor patient suffering of congestive heart failure; in [22], the author proposes a system to monitor cardiac activities using wireless wearable sensors and an *Android* device, taking into consideration cost and ease of use; *GiraffPlus* [23] project belongs to the Programme for Research and Technological Development in the Seventh Framework (2007-2013), which proposes a monitoring system solution based on environmental wireless sensors and a robot recognising user's activities and physiological parameters; in [24], the author proposes a work to monitor in mobility patients using *6LoWPAN* (*IPv6 over Low-Power Wireless Personal Area Networks*) devices in a hospital; in [25], the author presents an *NFC* and *RFID* based telemonitoring mobile solution to monitor blood pressure and heart rate; in [26], the author shows a mobility medication control system exploiting *RFID* technology.

With the growth of information coming from numerous sensors, the management of health data is crucial. Literature presents two interesting and innovative approaches to health data management: in [27], the author offers an engine for querying data from heterogeneous sources; in [28], the author presents a cloud computing system based on data mining procedures to correlate health and lifestyle.

All the presented works are examples of *IoT* technology applied to the health field. Unfortunately, virtually all have been designed for, and are strictly tied to, specific proprietary hardware or devices and technological communication standards, creating quite strong constraints on the choice of devices to be used for monitoring. To overcome such limitation and enable devices to be chosen based on their functionalities, cost and performance, regardless of purely technical and compatibility aspects, a system is needed that allows different devices using different communication standards to work together at the same time.

Current literature studies apply the social concept to *IoT* in two different ways: devices interacting directly with humans, following the example of Social Networks (*SIoT* between devices and humans) [29] and devices that interact amongst themselves in a *Social IoT* way (*SIoT* between devices). This twofold interpretation of *SIoT* has been described in [30], which also provides some examples of interactions between devices. In [31], the author extends and formalizes the differences between these two scenarios by defining the integration of devices within Social Networks as the *Web of Things* (*WOT*) and direct interactions between devices as the *SIoT* proper. They also detail the state of the art of the underlying technologies and apply these concepts to Wireless Sensor Networks. In [32], the author presents a solution and a prototype able to establish communications between two or more heterogeneous devices by exploiting their proximity (the communication takes place only when devices are near each other). Once again, most of the *SIoT* approaches proposed in the literature are designed for specific proprietary hardware or devices.

With regard to the current state of the art, herein we present a new home automation interoperability architecture based on the "friend" concept applied to services and devices. The greatest advantage offered by the system is that it is designed to integrate new devices, home automation systems or sensor networks, even if they communicate through different, incompatible native protocols. Exploiting the *SIoT* paradigm, devices and services communicate and share values and states with each other directly, without the need for intermediary components such as a database.

One direct consequence of Social paradigm approach is the advantage of being able, as needed or convenient, to freely replace any device or appliance of a specific brand with another brand's device without interfering with the functioning of the entire system. It moreover allows for the possibility of supporting the

network application layers implementing the *SIoT* functions for different purposes, as well.

4. Socialize Project

The *SOCIALIZE* (Service Oriented Assisted Living Infrastructure) project aims to meet the needs of today's (and future) active ageing. The projects aims to create a complete technological Web solution which integrates into a single platform a multiplicity of services designed specifically for the senior population and their caregivers (secondary end users). These services aim to strengthen social relations among the elderly, and with their relatives, who often live far away, and to create a sense of security and autonomy at home. The distance separating relatives from and older person living alone at home represents a potential health and security risk. In fact, the elderly are more subject to chronic diseases and unexpected critical health events. For these reasons, it is crucial to be able to monitor their health status and activities in order to identify any possible worsening of their conditions and avoid dangerous situations.

Although the *SOCIALIZE* Project is currently aimed at active ageing people with a certain degree of knowledge about the use of technological equipment, the ultimate goal is to make the platform accessible to all elderly people, including and especially, the large proportion of the older population that nowadays has great difficulty accessing and using *ICT*. For this reason, the user interfaces have been given rather simplified designs to make them easy to use, thereby permitting the inclusion of the largest number of elderly as possible.

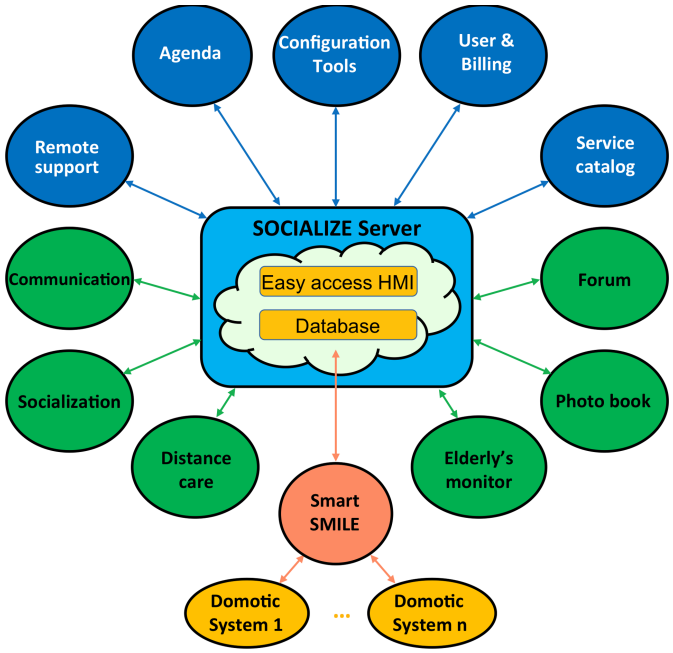


Figure 1: SOCIALIZE services

The services supplied by the project (Figure 1) can be divided in three categories:

- *Socialization* services: which aim to improve older people's social interactions. These include: (a) *communications*: enables exchanging messages and making audio / video call between the members of the *SOCIALIZE* community; (b) *socialization*: a dedicated social network between *SOCIALIZE* members; (c) *forum*: an advanced application for discussion of topics of general and personal interest; it can also highlight the most interesting discussions; (e) *photo book*: for storing and organizing photos and videos; (f) distance care: can follow the member's movements and monitor daily activities, and even warn of any threats; (g) *monitor*: analyzes user activities, as well as personal and environmental parameters; the results of this monitoring can be supervised by doctors, relatives and / or caregivers as is deemed appropriate with regard, for example, to the cognitive, health status and wishes of the elderly;
- *Cross services*: vertical applications common to all socialization services. The cross services currently implemented are: (a) *configuration tool*: provides for configuration of the system and services; (b) *user & billing*: manages *SOCIALIZE* platform users and their access credentials; (c) *remote support*: which provides help on using the *SOCIALIZE* platform or, more generally, computers or the device used; (d) *agenda*: collects all the events and date/time-related activities for a given user; (e) *service catalogue*: manages the available services, enabling them to communicate through the use of an enterprise bus system;
- *Cloud services*: it allows to the different services provided by the platform to communicate by sharing data and functionalities and it hosts the database containing data coming from the different Socialize service components. It consists of: (a) *easy access Human Machine Interface (HMI),* a web application with public and private content that acts as the entry point to the *Socialize* platform; (b) *database*, containing all the information related to users and services.

5. SmartSMILE framework

5.1. Concepts and use cases

The elderly monitor service is part of the *SOCIALIZE* platform. Its aim is, on one hand, to monitor elderly people's activities and physiological parameters to display processed data to caregivers, doctors, and relatives, to notify them of suspicious values, and on the other, to combine device functions to improve the individuals' life at home, increasing their autonomy and making them feel "protected" and "safe". To these ends, the *SmartSMILE* (Socialize MIddLeware Elderly) framework was created. *SmartSMILE* is a monitoring and interoperability *SIoT* ready software infrastructure for the socialization between devices. It is based on *SOA Web Services* and *XML* technologies to create intelligent domestic environments whose components collaborate to provide for the integration and interoperability of devices and household utilities belonging to different, incompatible native technologies. In fact, *SmartSMILE* creates an abstraction layer that overcomes the limitations of device compatibility, thereby enabling both the monitoring and composition services to work irrespective of the technologies of individual devices.

Figure 2 shows a monitoring use case. When a device changes its state, the new value is collected in a repository. When needed, all stored values are retrieved to be displayed to caregivers, doctors and family members. A device state change can in general occur for three main reasons:

- Direct interaction: the user physically acts on the device to change its status (e.g. pressing a button);

- Indirect interaction: the user acts on the device without touching it (e.g. through a smartphone);
- Involuntary interaction: the user does not interact with the device, but its state changes as the result of some event (e.g. a PIR sensor detects a presence in the environment).

Figure 3 shows a use case related to activation of the interoperability function. When a device changes its state, *SmartSMILE* verifies the condition to activate interoperability between devices. To do this, *SmartSMILE* (Figure 3) activates a *friendship manager* that searches for a friend relation that involves the device. If one is found, the other device is identified and the corresponding function is automatically activated. A device function is a command that any given smart device can execute. For example, the functions for a lamp can be *switch on*, *switch off* and *getStatus*. The *friendship manager* is a special component of *SmartSMILE* that defines which devices can interact with each other, and how, to create a composition of functions. As friends, two or more devices are trusted and are thus enabled to access each other's private data and functions, thereby providing interoperability.

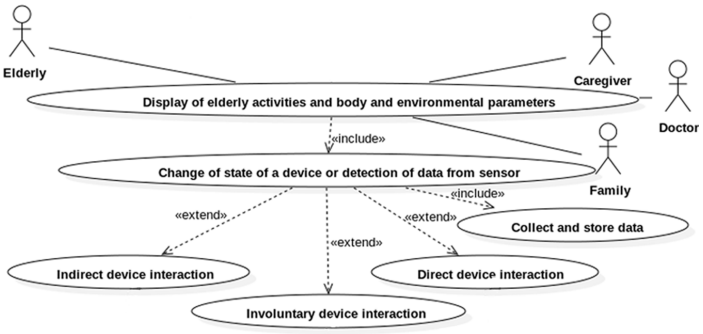


Figure 2: Monitoring use case

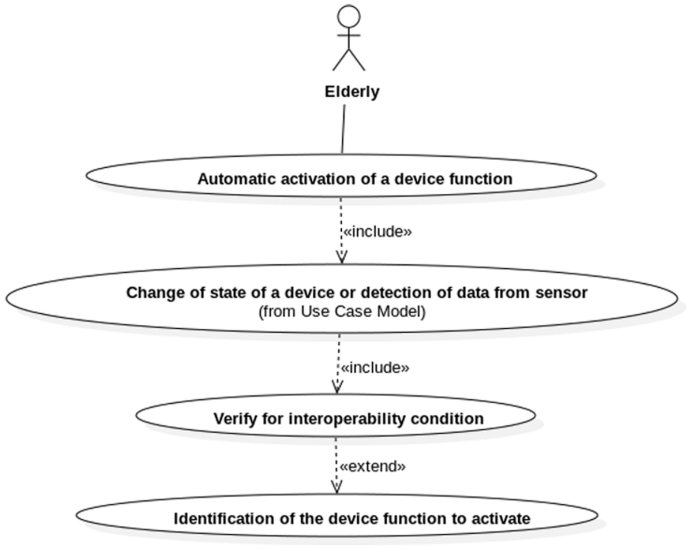


Figure 3: Interoperability use case

5.2. Architectural description

The *SmartSMILE* architecture (Figure 4) is made up of a set of sub-modules, called gateways. Each gateway interfaces with specific domotic systems expressly installed inside the home. The gateway serves the function of managing its own set of devices, integrating and interfacing them with the platform. To uniformize devices belonging to different gateways, *SmartSMILE* uses a high-level ad hoc language, called *SmartSmileML*, to describe devices in a technology-independent fashion. Such description includes their main characteristics, and the functions that they can execute, along with their input and output parameters and defined data types. Moreover *SmartSmileML* describes the possible events (e.g. device state change) specific to the supported domotic technologies for processing. To describe devices, functions and events, *SmartSmileML* uses an XML based language. Its functioning does not require prior definition of taxonomies or classifications for devices and events. By avoiding the use of pre-defined devices, the system is able to manage any kind of device, exploiting all the available functions that are automatically identified during its creation.

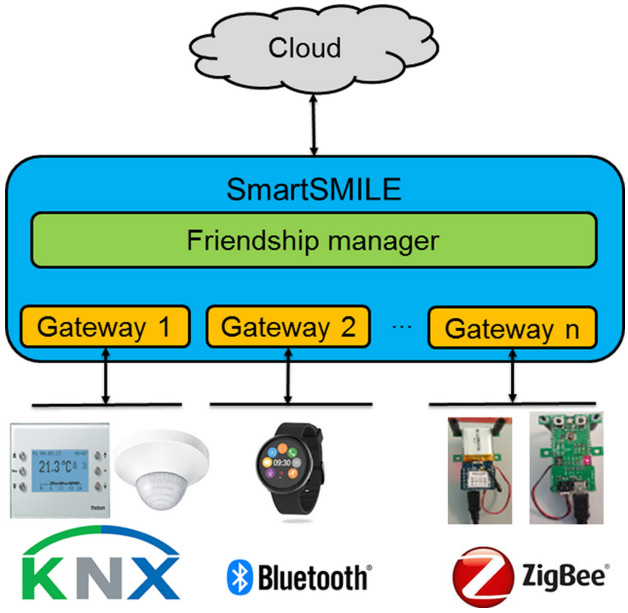


Figure 4: SmartSMILE architecture

In order for the system to work, each gateway has to implement some basic features; these are:

- Creation of the high-level description of devices belonging to its network: This operation enables having a unique view of, and access to all the devices belonging to *SmartSMILE*. In this way, *SmartSMILE* can use a unified way to interact with them. For this to work, the gateway must be able to map the device descriptions with the corresponding real devices.
- Capture of events that occur in the automated home network and translation into the *SmartSmileML* language. This allows for monitoring events in the domotic bus and enables *SmartSMILE* to react accordingly. The obtained information are translated into the *SmartSmileML* and are sent to the cloud server for collection and querying as needed.
- Translation of a *SmartSmileML* function description into the corresponding domotic command. This enables *SmartSMILE* to invoke commands in devices. To do this the system must know the corresponding gateway managing each device.

5.3. SmartSmileML language

SmartSmileML consists of two main formalisms: *SmartSMILE Device* and *SmartSMILE Message.*

SmartSMILE Device

SmartSMILE Device: defines devices and their functions. In particular, it describes the characteristics of each device and the processes by which it interacts with other *SmartSMILE Devices*. The XML description includes the following tags:

- *device*: the *root* element that is the *parent* of all other elements. The main attributes contained in this tag are: (a) *description*: a free text field; (b) *id*: the identifier of the device within the *SmartSMILE* framework; such identifier must be clearly unique; (c) *serialNumber*: a device identifier in the domotic technology; (d) *tech:* the domotic technology employed by the device; this allows for immediate assignment of the device to the proper gateway; (e) *type*: the kind of device; this can be a free text field or contain pre-defined values;
- *service*: *child* of the *device* tag. There is a *service* tag for each device function,. The main attributes are: (a) *name*: the name used in the relevant domotic technology to call that function; (b) *output*: the *SmartSMILE* data type; this attribute is defined, only if the function returns a value;
- *input*: *child* of the *service* tag. There is an *input* tag for each input parameter that the function requires,. The main attributes are: (a) *name*: a label for the input; (b) *type*: the *SmartSMILE* data type;
- *allowed*: *child* of the *input* tag. There is a tag for each possible value that that input value can take. If no allowed tags are defined, it means that any value is allowed for the declared *SmartSMILE* data type (e.g. if the type is *int,* and no *allowed* tags are defined, the input can be any integer value. The main attribute is *value*: the allowed value;
- *linkedService*: *child* of *service* tag. If defined, it describes friendship relations with other devices. The main attributes are: (a) *id:* the id used in *SmartSMILE* to identify the friend device; (b) *service*: the name of the function of the friend device to invoke to create compositions;
- *linkedInput*: *child* of *linkedService*. There is a *linkedInput* tag for each input of the friendly device function to invoke,. The main attributes are: (a) *from*: the name of the input to the described function whose value is needed in order to be shared with the friend device function named in the *linkedService* tag; (b) *to*: the name of the input of the friendly device function whose value has to be assigned using the shared one; (c) *value*: if *from* attribute is not defined, this defines a fixed value for the *to* input name.

SmartSMILE Message

SmartSMILE Message formally describes events, commands and responses. The XML description establishes the following tags:

- *message*: the *parent* tag. The main attributes contained in this tag are: (a) *message*: the name of the function involved in the message; (b) *messageType*: defines the scope of the message: the possible options are: *COMMAND*, when the message is addressed to a gateway for translation into a domotic message and written to the domotic bus for execution. In this case, the *message* attribute is related to the function to activate; UPDATE, when an event related to a device change of state is captured on the domotic bus, it is transmitted to the *Friendship manager* to check whether the conditions for interoperability are satisfied. In this case, the *message* attribute regards the device function that can be correlated to the change of state; (c) *senderId*: if the message is of type *UPDATE*, the id of the *SmartSMILE* device involved in the message; (d) *receiverId*: if the message is of type *COMMAND,* the id of the *SmartSMILE* device involved in the message;
- *input*: *child* of the *message* tag. This describes the values that are involved in the message. The main attributes are: (a) *name:* the name of the value; (b) *type*: the *SmartSMILE* data type; (c) *value*: the new value.

5.4. SmartSMILE SIoT features

Using *SmartSmileML*, it is possible to achieve interoperability among services and devices by taking advantage of a single means of communication and data sharing. Thus, it is possible to achieve a social network of intelligent objects, whereby devices are tied to each other by social relationships in order to collaborate and provide a common *SIoT* application. When a message passes through a domotic bus, it is captured by its gateway. The gateway identifies the correspondence between the device that sent the message and the associated description in *SmartSMILE Device*. Using the information in the device description, the captured event is translated into the *SmartSmileML* language, inserted into *SmartSMILE Message* and forwarded to the *Friendship manager* to check for the existence of relations, and finds the device and its function that can achieve the necessary interoperability. To verify and activate the friendship, the *Friendship manager* checks the *linkedService* and *linkedInput* tags to verify the relation between the involved devices, and if it exists, it shares input and output values between them.

By way of example, let us assume we have two devices belonging to different domotic systems whose functions have to be combined as shown in Figure 5. The first device is a "wall switch" employing the *KNX* [33] technology. It has a function named SET_STATUS (the content of the tag *service name*), whose input name *status* is an *OnOff* data type. The values of *OnOff* can be *on* or *off*. The second device is a "lamp", which instead uses *UPnP* [34] technology. It has two functions: *getState* and *setPower*. Both use a value of type *OnOff* as output and input respectively. The *device description* of the function *SET_STATUS* of the wall switch has a special tag, named *linkedService*. This tag permits the *Friendship manager* to establish a friend relation between the two devices and to create a composition of functions. When the *Friendship manager* receives the event that the wall switch is set to *on* (Figure 6), it 'knows' that it has to invoke the function *setPower* of the device with *id = 3* (the lamp). The content of the *linkedInput* tag specifies that the wall switch shares the *status* value with the lamp. In the end the *Friendship Manager* invokes the lamp's *setPower* function, which will have the value *on* within the input named *power* (Figure 7).

The device descriptions are for the most part created automatically by the gateway. The complexity involved in this depends on the standard to be managed. For example, in *UPnP* the creation is completely automated because the devices present and

describe themselves each time they enter the network. For *KNX*, the descriptions are made by parsing a configuration file created by the configuration software named *ETS* [35]. In some cases, automating the process is not possible and it is necessary to write them manually.

```
<device description="wall switch" id="1"
  serialNumber="123" tech="KNX" type="switch">
  <service name="SET_STATUS">
    <input name="status" type="OnOff">
      <allowed value="on" />
      <allowed value="off" />
    </input>
    <linkedService id="3" service="setPower">
      <linkedInput from="status" to="power" />
    </linkedService>
  </service>
</device>

<device description="energetic saving lamp" id="3"
  serialNumber="456" tech="UPNP" type="lamp">
  <service output="OnOff" name="getState"  />
  <service name="setPower">
    <input name="power" type="OnOff">
      <allowed value="on" />
      <allowed value="off" />
    </input>
  </service>
</device>
```

Figure 5: Two device descriptions in smartSMILE Device

```
<message message=" SET_STATUS " messageType="UPDATE"
senderId="1">
   <input name="status" type="OnOff" value="on" />
</message>
```

Figure 6: The description of an event in SmartSMILE Message

```
<message message="setPower" messageType="COMMAND"
 receiverId="3">
   <input name="power" type="OnOff" value="on"/>
</message>
```

Figure 7: Command representation in SmartSMILE Message

6. Implementation

SmartSMILE exploits W3C-recommended standard Web technologies. It exposes services using *Web Services*, *SOA* and *XML* technologies. The advantage of using *W3C* standard solutions is that they ensure that the developed applications are fully compatible with other standards-based software and are not tied to any particular software system, programming language or computer architecture.

illustrates the deployment of the *SmartSMILE* component used in the *SOCIALIZE* Project. *SmartSMILE* is connected on one side to the Cloud server, where the values coming from devices are stored and recovered for any further analysis, and on the other side, to the wired or wireless domotic buses. Through the intervention of *SmartSmileML*, communications between the Cloud server and *SmartSMILE* are performed using high-level messages as shown in Figure 6. In this way, the cloud does not need to 'know' how the different domotic technologies work in order to understand messages – it just needs to be able to "speak" *SmartSmileML*. In addition to collecting values, the *Cloud* provides a web server environment and a web application to show each user's data to the caregivers. To match the received values from *SmartSMILE* with the respective user, the *Cloud* is previously configured associating devices to consumers.

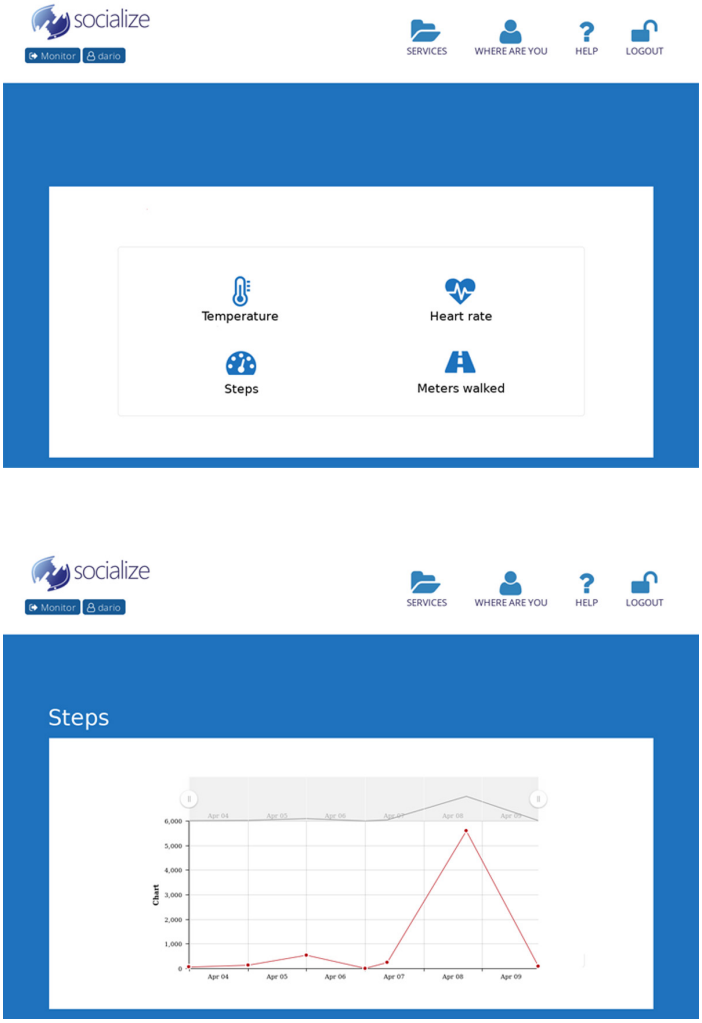


Figure 9: Example of "Steps" page of the web application

The different domotic technologies used in the project were: KNX, UPnP, MyHome [36], ZigBee and Bluetooth [37] KNX gateway. Table 1 shows the implemented classes for each developed *SmartSMILE* gateway. A brief description of each class is provided in the following subsections.

6.1. KNX gateway

To integrate *KNX* into *SmartSMILE*, the gateway exploits the API provided by the *Calimero 2* [38] libraries to connect to, capture and write KNX messages to and from the domotic bus. At startup time the gateway connects to the bus and initializes the required structures for translation of the data types and values used in *SmartSmileML* to KNX *Datapoint* identifiers and vice-versa. Then, the gateway loads the *ETS* (KNX Engineering Tool

Software) configuration file that contains information about different *KNX* devices connected to the KNX bus, such as their description, their application programs, and their functionalities. It then implements a listener on the KNX bus to manage events. Captured message are translated into *SmartSMILE Message* and can be of two types: a reply to a read request for a device state, or detection of an event that implies a device change of state that can trigger a composition of functions. To invoke functions, *SmartSMILE Message* are translated into KNX packages and written to the bus.

Table 1:Used Java Classes to implement gateways

Gateway	*Java Classes*
KNX	*KNX* *KNXListener* *KNXServiceMapping* *SSServiceData* *KNXServiceData* *SSParamterMapping* *KNXDriverException* *SSTypeValueXlator* *SSTypeValue*
UPnP	*UPnPManager* *UPnPManagerPoint* *UPnP2SmartSmile*
MyHome	*BTicino* *BTicinoDeviceData* *BTicinoDeviceManager* *BTicinoMsgsVerifier* *BticinoServiceData* *BticinoServiceManager* *BTicinoDeviceManager* *BTicinoDimensionValuesMapper* *SSServiceData*
SmartSmileML	*SmartSmileMLGateway*

To implement the gateway, 9 Java classes have been developed (Figure 10):

- *KNX*: this is the main class that acts as a communications bridge between *KNX* and *SmartSMILE*. It checks all the parameters of configuration files and establishes connections with the *KNX* bus via the *KNX/IP* gateway device. This makes it possible to read and write to the *KNX* bus. Moreover, it creates abstract instances of the *KNX* devices in *SmartSMILE* and maps them to the actual ones, so they can be used by the platform. The class also implements the reading and writing operations on the bus, and their conversion from *KNX* to *SmartSmileML* format and vice versa, including the values of the different data types.
- *KNXListener*: this class implements the functions and interactions with the *KNX* bus. It creates and manages the data flow to and from the *KNX* bus by providing: (a) synchronous reading of data (when the value of a datapoint is explicitly requested through a group address); (b) asynchronous reading (the system furnishes the value of a datapoint through the group address without any explicit request, such as a thermostat that periodically reports the temperature, or notification of a user activating a device); (c) writing for activation of functions. Using the configuration data of *KNX* devices in *SmartSMILE*, it is possible to associate the correct coding/decoding type of data (*DTP*) to each *group address*.
- *KNXServiceMapping*: defines the mapping between the *KNX group address* and the associated *KNX* data type.
- *SSServiceData*: once both the *individual address* of the device and the corresponding *KNX group address* are known, this Java Class obtains the appropriate *SmartSMILE* function.
- *KNXServiceData*: takes as input the *SmartSMILE* function and finds both the *individual address* and the *KNX group address.*
- *SSParamterMapping*: represents the data structure that allows mapping a *SmartSMILE Message* to the available input and output parameters.
- *KNXDriverException*: defines the exceptions generated during interactions with *KNX*.
- *SSTypeValueXlator*: implements the functionalities to convert a *KNX* value into a *SmartSMILE* value and vice versa, according to the respective data types. By way of example: a) the *KNX* value "on" can be converted to a *SmartSMILE* "true" Boolean value; b) the *SmartSMILE* percentage scale has a range of 0 to 100, while in *KNX* it ranges from 0 to 255; it is thus necessary to convert the values between the two scales.

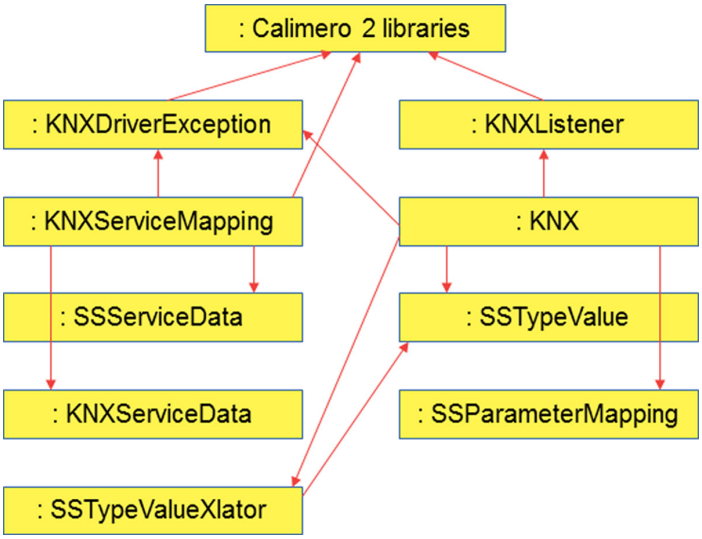


Figure 10: Developed classes for KNX gateway

When a function is invoked using the *SmartSMILE Message* formalism, the *KNX* class searches for the corresponding *KNX* service retrieved via the *KNXServiceMapping* class (using the *getKNXService* method). The value initially expressed in *SmartSMILE* format, is taken from the *KNXServiceData* class using the *GetValue* method. It is converted to the *KNX* value using the *SSTypeValueXlator* class with the get *getKNXParameterValue*

method. The data are then written to the *KNX* bus using the *KNXListener* class. In the end, a reply indicating the execution of the command is sent back, once again using the *KNXServiceMapping* class (method *getSSService*) in order to generate confirmation of execution. (Figure 11).

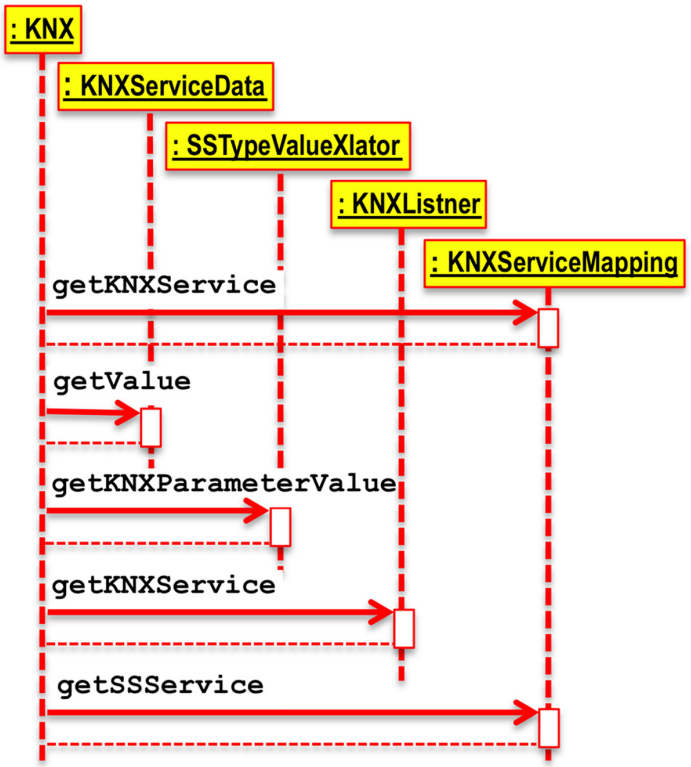


Figure 11: Write operation on KNX gateway

These operating principles of the *KNX Gateway* have also applied to the other gateways implemented in the project. In particular, regarding the development of the *UPnP* and *MyHome* Gateways.

6.2. UPnP gateway

nP technology, the gateway exploits the API provided by the *Cyberlink* and *Cidero* libraries [39]. At startup time, the gateway initializes the *ManagerPoint*, which is a *UPnP listener* that enables adding, removing and executing functions on devices. The gateway initializes the data structures for the translation from *SmartSmileML* to the corresponding *UPnP* values and types, and vice-versa. When a new device enters the *UPnP* network, it announces itself and provides a description of itself together with a list of functions that it can offer. This information is represented using *XML* structures that the *ManagerPoint* parses to create the *SmartSMILE Devices*.

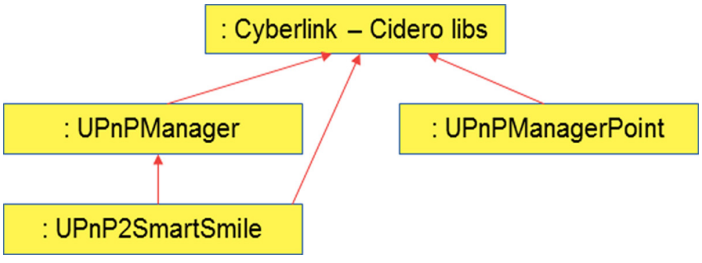


Figure 12: Developed classes for UPnP gateway

Three Java classes have been developed to implement the gateway (Figure 12):

- *UPnPManager*: this is the main class that acts as a communication bridge between *UPnP* and *SmartSMILE*. It checks all the parameters of configuration files and establishes interactive connections with *UPnP*. Moreover, it creates abstract instances of the *UPnP* devices in *SmartSMILE* and maps them to the actual ones. The class also implements the reading and writing operations on the *UpnP* bus and their conversion from *UPnP* to *SmartSmileML* format and vice versa, including the values of the different data types.
- *UPnPManagerPoint*: this class implements the functions of and interactions with *UPnP*. It creates and manages the data flow to and from the *UpnP* system by providing the synchronous and asynchronous functions for reading and writing to the bus.
- *UPnP2SmartSmile*: implements all the data structures and methods to convert descriptions, commands and values of *UPnP* into *SmartSmileML,* and vice-versa.

6.3. MyHome gateway

library was used, though modified to add the authentication functionalities offered by the *AndroidBetaOwnAuthentication* library. This library makes it possible to manage communications through an *OpenWebNet* (*OWN*) server by formatting messages according to the *OWN* syntax. It moreover serves to open sessions, creating special threads, and send and receive reply messages as well as syntactic and communication errors. At startup, the gateway opens communications with the *OpenWebNet* server that interfaces with the *MyHome* technology. The server works using two sessions: monitor and command. During the monitoring session the gateway starts a thread that remains in execution to asynchronously intercept the events that it reads from the bus. Messages are parsed by an *LL(1)* [41] grammar. During the command session, instead, the gateway establishes a connection with the server when it has to invoke a function. Unfortunately, auto discovery and auto configuration of *SmartSMILE Devices* is not available because the technology does not offer any device discover process. Hence, it is necessary to create the *SmartSMILE Devices* manually.

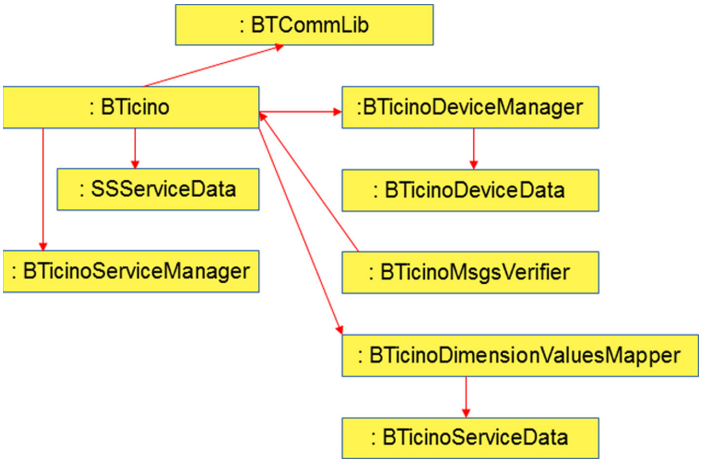


Figure 13: Deployed classes for MyHome gateway

To implement the gateway, 8 Java classes were developed (Figure 13):

- *BTicino*: this is the main class that manages devices and creates all the structures needed to interact with *MyHome* systems. It creates the two threads needed to interface with the monitoring and command sessions.
- *BTicinoDeviceData*: contains data and information related to a *BTicino* device for interactions.
- *BTicinoDeviceManager*: manages *BTicinoDeviceData*, organizing them in a Hash table.
- *BTicinoMsgsVerifier*: checks messages from the monitoring session.
- *BTicinoServiceData*: contains data and information related to *BTicino* messages.
- *BticinoServiceManager*: organizes BTicinoServiceData in a Hash table.
- *BTicinoDimensionValuesMapper*: translates quantities from *SmartSmileML* to *OWN* and vice-versa.
- *SSServiceData*: contains data and information useful for executing a *SmartSmile Message*.

6.4. *SmartSmileML* gateway

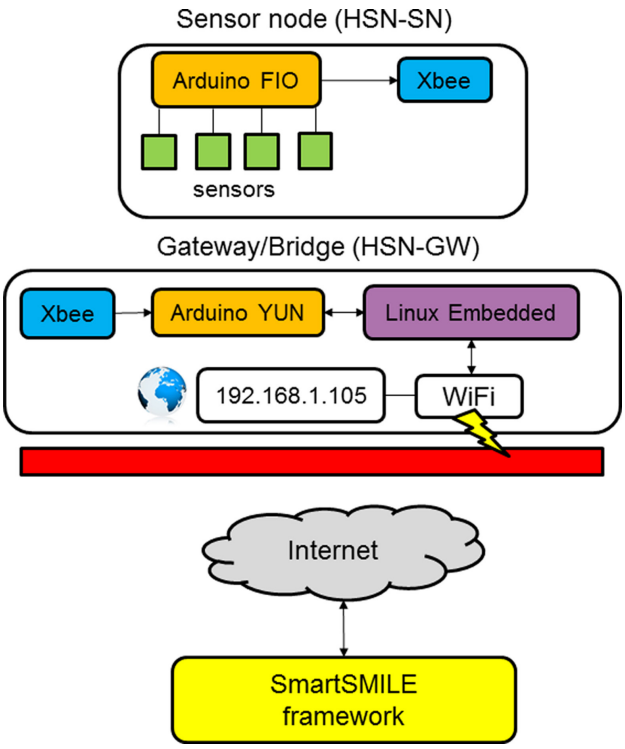
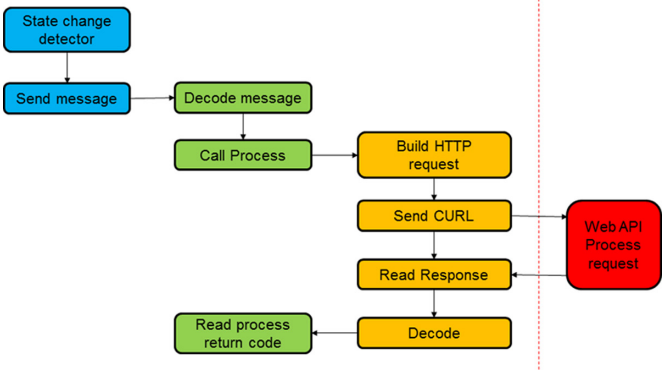


Figure 14: ZigBee architecture

The *ZigBee* architecture (Figure 14) used in the project is fully IEEE 802.15.4 and ZigBee 2007/PRO compliant. The transmitting protocol is ZigBee HA profile. It is composed by two main elements:

- *Smart node* (HSN-SN): this is a device made up of a sensor, an *XBee S2* radio module configured as a *ZigBee* router and a node. The core of the node is an *Arduino FIO* board with one or more sensors connected to its I/O ports (input/output lines, either digital or analog). Each sensor node receives broadcast messages from the *coordinator* (bridge) and sends events only to the *coordinator* in a so-called *star* network topology. Events generated from sensor nodes are changes in the state of a sensor, power up or reset, low battery and heartbeat (a timed event indicating the node is alive).

A simple character-oriented communications protocol has been implemented on the smart node for bi-directional communication, network management and control of the input/output lines between smart nodes and the gateway/bridge. Events notification messages are forwarded directly, while the incoming commands require some tweaks on the server side. Although the communication protocol is able to manage bidirectional messages (from sensor node to gateway and vice versa), only one way is used (Figure 15).



Bluetooth devices in the project are interfaced using an *Android Smartphone* app (Figure 16).

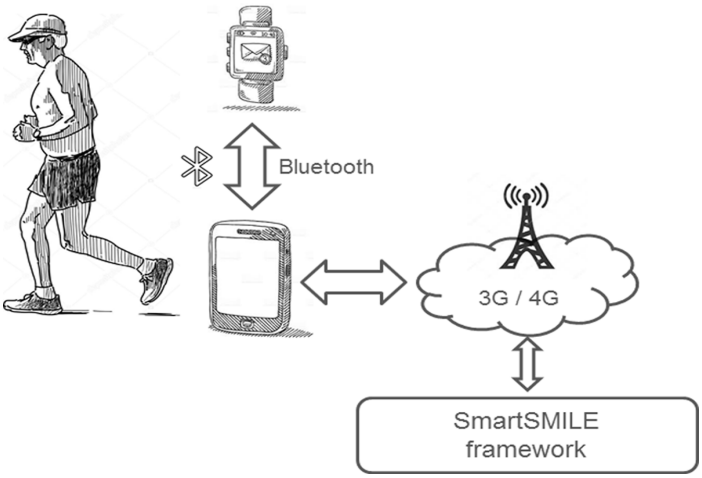


Figure 16: Bluetooth device connected to a Smartphone

Both the *ZigBee* Gateway and the *Android* app for *Bluetooth* are specifically designed to communicate through the *SmartSmileML* language. This provides *SmartSMILE* with a native *SmartSmileML* gateway, which has the task of setting up a connection as a *TCP* or *UDP* socket or a *REST* service to which devices can send messages. Incoming messages are directly forwarded to the *SmartSMILE* core for analysis by the *Friendship manager* and then forwarded on to the cloud server. Since this gateway does not require any special implementations to convert data, values and functions, it was sufficient to include all the features described in a single class called *SmartSmileMLGateway*.

In order to make devices able to 'speak' *SmartSMILEML*, device developers can refer to a document describing how *SmartSmileML* works and how to use it.

When a device becomes active, it sends a message in which it describes itself using a *SmartSMILE Device* description, as shown in Figure 5. In this case, the attribute *tech* is *SmartSMILE*. The *id* has been left blank because *SmartSMILE* will automatically fill it in. The assigned *id* will be used by the device for its communications. When the device is deactivated, it sends a special *SmartSMILE Message* of the type *REMOVE* containing its *id* (Figure 17).

```
<message id="90A2DAF501A4.ROOM105.NODE101"
messageType="REMOVE" />
```

Figure 17: SmartSMILE Message of type REMOVE

Leveraging the *IoT* capabilities of *SmartSMILE*, we ran a pre-test of the proposed system using remote devices during the development stage. A *SmartSMILE* server was deployed in ... with various devices employing *KNX*, *UPnP* and *MyHome* technologies, including a thermostat, motion detectors, gas and smoke detectors, electrical load control of instantaneous consumption, power actuators, a cooking gas electromagnetic valve and smart TV. All the appliances and devices are part of a demonstration installation that simulates an actual home. *ZigBee* environmental sensors, including temperature and humidity, were fitted at the *SUPSI University* in *Switzerland* in *Manno*. A *Bluetooth* Smartwatch able to capture users' heart rates, number of steps and sleep duration was set up at *Spring Techno* in *Bremen Germany* (Figure 18).



Figure 18: Smartwatch and ZigBee sensors

7. Field trials

The test and validation processes of the *SOCIALIZE* platform were performed by the end-user partners in the project.

In Switzerland it has been tested in a retirement home, while the Italian test involved two regions (Tuscany and Sicily), making sure in each case to involve different types of target users in the test:

- Care-dependent elderly people in residential facilities;
- Partially autonomous elderly people assisted at their own residence/homes;
- Autonomous elders who are not beneficiaries of social care or healthcare assistance and participate in associational activities.

Regarding the elderly monitoring service, the test was conducted inside a residential facility in Tuscany. The aim was to collect data from sensors and ensure that they were understandable, and that they were able to constantly follow and track the evolution of those parameters deemed significant for the health and safety of the older persons involved.

A technologically enriched environment was recreated in this residence by assembling and recreating within the pilot site the same*IoT* device configuration already used in the earlier design stage.

The testing period was two weeks. *SmartSMILE* was installed on a computer located inside the residence and connected to the Internet through an *ADSL* connection.

The devices installed were (Table 2): a wireless ambient temperature and humidity sensor in the cafeteria (*ZigBee* technology), a temperature and humidity sensor as well as a SmartTV and Media Server (*UPnP* technology) in the dining room, two PIR wired presence sensors, one in a bedroom and one in the kitchen, which was also equipped with gas electrovalve actuator, push button, gas, water, smoke sensors (*KNX* technology) and electrical load control (*MyHome* tecnology), and finally two SmartWatches worn respectively by two guests (*Bluetooth* technology). Almost all *IoT* devices used are powered by a mains power source, only the *ZigBee* temperature and humidity monitoring devices could have a battery life problem.

Table 2: Devices in field tials

Device	Technology	Location
Ambient temperature and humidity sensor	ZigBee: wireless bus with plug and play feature. Battery power supply	Cafeteria Dining room
Smart TV and Media Server	UPnP: wireless WiFi bus. Mains electricity supply	Dining room
PIR presence sensor	KNX: wired bus. Mains electricity supply	Bedroom Kitchen
Gas, water and smoke sensors Gas electrovalve and push button Power actuator	KNX: wired bus. Mains electricity supply	Kitchen
Electrical load control	BTicino: Wired bus. Mains electicity supply	Kitchen
Smartwatches	Bluetooth: wireless bus. Rechargeable battery power supply	Wearable device

They have 2 AA (LR-03) batteries and their measured energy consumption is 35mA in transmission and 2.1uA in sleep mode. The run time is 1 year based on sending data per 20 min. Both *SmartSMILE* framework and the corresponding gateways to manage the domotic technologies run on a *Raspberry Pi 3 Model B+* [42].

The data collected by these heterogeneous home automation devices was converted, standardized and normalized by *SmartSMILE*. The data were then sent to a database managed in Cloud Computing mode residing at the Digimat company, in Matera, Italy.

The results of the experimentation made it possible to verify the correct functioning of the entire system. In particular, through the use of dedicated Web interfaces it was verified that data were stored correctly. These interfaces have enabled simple, intuitive aggregation of the data over fixed periods of time and their display in clear, easily understandable graphics.

In fact, the information is displayed as simply as possible, with icons corresponding to the requested service (for example Physical Activity or Heart Rate monitoring). The end user needs only click the corresponding icon to show a graphical representation of the acquired data with the possibility to adjust the timeframe.

The proper functioning of the collaboration and interoperability between the devices has also been verified, as these features are necessary for the implementation of a *SIoT* based on the *SmartSMILE* infrastructure.

The tested monitoring activities included:

- Temperature control inside the home environment. This has led to the discovery that people who tend to keep the room temperature either too hot or too cold are subject to falling ill more frequently;
- Supervision of elderly people who do not get enough physical activity with respect to doctors' recommendations;
- The feeling of safety older people experiences with regard to the hazards of fire, flooding and gas leaks that may occur within the environment.

The strength of the *SIoT*-ready infrastructure is highlighted by the following composite functions:

- Following the voluntary provocation of conditions of flooding, gas leaks and smoke in the kitchen using dedicated *KNX* devices (Figure 19), when a hazardous event was recognized, the corresponding alarm signals from the *UPnP* media server were promptly displayed on the television, together with the simultaneous activation of an acoustic alarm.
- When one or more appliances requiring high power consumption were turned on at the same time (specifically the microwave oven, the traditional oven, the boiler and the washing machine), the electrical load monitoring system of detected an overload. Power actuators were deactivated according to a preset priority until the situation returned to normal. When the load reduced because one of the still functioning household appliances finished, the load control detected this change and reactivated the appliances that were previously stopped. The functionalities were implemented using *MyHome* electrical load control of instantaneous

);



Figure 20: Electrical load control and power actuator

witching on the cooking gas requires pressing a button. After a pre-set period of time, an acoustic and luminous alert was activated, and an appropriate message displayed on the TV to remind the user that to continue cooking the button had to be pressed again. This prevented the user forgetting food on the cooker. The combined timer and smart TV functionalities were composed using the *KNX* electromagnetic cooking gas valve (Figure 21).

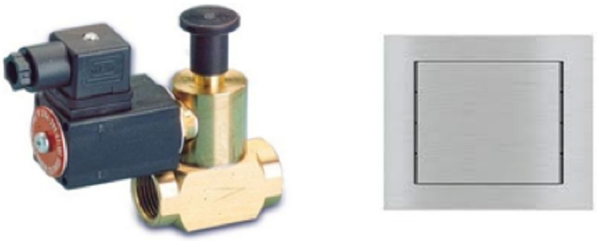


Figure 21: Gas electrovalve and push button

The most significant anomalies identified were long periods of no information coming from both SmartWatches. For the first incident, the reason was that the person took off the watch for 5 nights consecutively, while the second was due to the fact that twice (3 hours and 5.5 hours, respectively) the other person did not notice that the device was no longer working because the battery needed recharging.

8. Conclusions

The paper presented *SmartSmile*, a framework that solves the main issues so that *SIoT* paradigms can find actual application. This software infrastructure was developed as an activity of the *Socialize* project, to address the problem of monitoring older

people within the living environment.

Thanks to the potential of *SmartSmileML language*, the framework implements an abstraction layer that hides the underlying, often proprietary technologies, of the integrated devices, leveling them out. In addition, *SmartSmileML* defines a strategy to define the "friendship" relation between devices permitting the sharing of information amongst them even if they are not natively interoperable. On the one hand, *SmartSmile* permits to users to choose the most suitable sensors/actuators for the specific case without concern for their compatibility with the devices or systems already installed and, on the other hand, it provides developers with a unique representation of the acquired data without the need to know the technical details of each single technology.

Summarizing, the most significant features that have allowed the creation of a *SIoT* environment supporting the services of device discovery, combination of device functions and reliability, are:

- Interoperability: the ability of the devices to interoperate with each other even if they belong to natively incompatible home automation systems, thereby permitting full interconnection and interaction amongst them.
- Service-to-service communications: the service provided via a device can share its input and output with the services of other “friends” to achieve a common objective.
- Scalability: in order to attain a certain goal, devices can collaborate with each other; moreover, when changes occur (new devices become available or some are missing), tasks can be re-distributed.

As a future work, *SmartSMILE* will allow to expose its own devices externally. Exploiting the results of the work described in [43], every device within the system will be reachable using an *IPv6* address, and socialization will be achieved between devices belonging to different *SmartSMILE* instances. In addition, the stored data could be made available to special machine learning algorithms that could launch alarms automatically, learning the user's habits.

SmartSMILE is released under the *GPL licence* [44]. Its code is shared on GitHub [45].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The *SOCIALIZE* project has been financed by the *European Commission* through the *Ambient Assisted Living Joint Programme* (AAL JP).

References

[1] V. Miori, D. Russo, "Improving life quality for the elderly through the Social Internet of Things (SIoT)" in IEEE Global Internet of Things Summit (GIoTS), 1-6, 2017. https://doi.org/10.1109/GIOTS.2017.8016215

[2] A. Börsch-Supan, M. Brandt, C. Hunkler, T. Kneip, "Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE)" International Journal of Epidemiology, 2013, https://doi.org/10.1093/ije/dyt088

[3] A. Hofman, D. E. Grobbee, P. De Jong, F. A. Van den Ouweland, "Determinants of disease and disability in the elderly: the Rotterdam Elderly Study" European journal of epidemiology 7.4, 1991. 403-422. https://doi.org/10.1007/BF00145007

[4] Kaare Christensen, Gabriele Doblhammer, Roland Rau, James W Vaupel, "Ageing populations: the challenges ahead", The Lancet, 374(9696), 2009. https://doi.org/10.1016/S0140-6736(09)61460-4.

[5] World Health Organization, World report on ageing and health, ISBN: 978 92 4 156504 2, 2015.

[6] M. Sánchez-Valle, M. Vinarás Abad, C. Llorente-Barroso, "Empowering the Elderly and Promoting Active Ageing Through the Internet: The Benefit of e-inclusion Programmes" Safe at Home with Assistive Technology, Springer, Cham, 95-108, 2017. https://doi.org/10.1007/978-3-319-42890-1_7

[7] J. Gubbi, R .Buyya, S. Marusic, M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions" Future Generation Computer Systems, Volume 29, Issue 7, 1645-1660, 2013. http://dx.doi.org/10.1016/j.future.2013.01.010.

[8] C. Christophorou, S. Kleanthous, D. Georgiadis, D. M. Cereghetti, P. Andreou, C. Wings, E. Christodoulou, G. Samaras, "ICT services for active ageing and independent living: identification and assessment" Healthcare technology letters, 3.3: 159, 2016. https://doi.org/10.1049/htl.2016.0031

[9] V. Miori, D. Russo, "Preventing Health Emergencies in An Unobtrusive Way", Internet of Things. User-Centric IoT. Springer, Cham, 242-247, 2015. https://doi.org/10.1007/978-3-319-19656-5_35

[10] L. Atzori, A. Iera, G. Morabito, M. Nitti, "The social internet of things (siot)-when social networks meet the internet of things: Concept, architecture and network characterization" Computer Networks 56(16), 3594-3608, 2012. https://doi.org/10.1016/j.comnet.2012.07.010

[11] L. Atzori, A. Iera, G. Morabito, "From "smart objects" to "social objects": The next evolutionary step of the internet of things" Communications Magazine, IEEE 52.1, 97-105, 2014. https://doi.org/10.1109/MCOM.2014.6710070

[12] http://www.socialize-project.eu/

[13] V. Miori, D. Russo, "Domotic Evolution towards the IoT" in 28th International Conference on Advanced Information Networking and Applications Workshops (WAINA 2014), Victoria, Canada, 809-814, 2014. https://doi.org/10.1109/WAINA.2014.128

[14] British Standards Institute. Industrial Automation Systems – Concepts and Rules for Enterprise Models, BS ISO 14258, London: British Standards Institute, 1998

[15] M. Ganzha, M. Paprzycki, W. Pawlowski, P. Szmeja, K. Wasielewska, "Semantic Technologies for the IoT - An Inter-IoT Perspective" in IEEE First International Conference on Internet-of-Things Design and Implementation (IoTDI), Berlin Germany, 271-276, 2016. https://doi.org/10.1109/IoTDI.2015.22

[16] A. Broering, S. Schmid, C. K. Schindhelm, A. Khelil, S. Kaebisch, D. Kramer, D. Le Phouc, J. Mitic, D. Anicic, E. Teniente, "Enabling IoT Ecosystems through Platform Interoperability" IEEE Software, vol. 34, no. 1, 54-61, 2017. https://doi.org/10.1109/MS.2017.2

[17] S. Kubler, J. Robert, A. Hefnawy, K. Främling, C. Cherifi, A. Bouras, "Open IoT Ecosystem for Sporting Event Management," IEEE Access, vol. 5, 7064-7079, 2017. https://doi.org/10.1109/ACCESS.2017.2692247

[18] S. Soursos, I. Podnar Zarko "symbIoTe: Symbiosis of Smart Objects Across IoT Environments, in Digitising the Industry" Internet of Things Connecting the Physical, Digital and Virtual Worlds, IERC Cluster Book, 303-307, 2016. https://doi.org/10.1109/EuCNC.2016.7561070

[19] Y. Guan, J. C. Vasquez, J. M. Guerrero, "An open virtual neighbourhood network to connect IoT infrastructures and smart objects — Vicinity: IoT enables interoperability as a service" in Global Internet of Things Summit (GIoTS), Geneva Switzerland, 1-6, 2017. https://doi.org/10.1109/GIOTS.2017.8016233

[20] S. Kubler, J. Robert, A. Hefnawy, C. Cherifi, A. Bouras, K. Främling, "IoT-based Smart Parking System for Sporting Event Management" in 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MOBIQUITOUS 2016), ACM, New York USA, 104-114, 2016. https://doi.org/10.1145/2994374.2994390

[21] J. H. Abawajy, M. M. Hassan, "Federated internet of things and cloud computing pervasive patient health monitoring system" IEEE Communications Magazine, 55(1), 48-53, 2017. https://doi.org/10.1109/MCOM.2017.1600374CM

[22] P. Kakria, N. K.Tripathi, P. Kitipawang, "A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors" International journal of telemedicine and applications, 8-19, 2015. https://doi.org/10.1155/2015/373474

[23] S. Coradeschi, A. Cesta, G. Cortellessa, L. Coraci, C. Galindo, J. Gonzalez, L. KarlssonA. Forsberg, S. Frennert, F. Furfari, et al., "GiraffPlus: a system for monitoring activities and physiological parameters and promoting social

interaction for elderly" Human-Computer Systems Interaction: Backgrounds and Applications 3, Springer, Cham, 261-271, 2014. https://doi.org/10.1007/978-3-319-08491-6_22
[24] M. S. Shahamabadi, B. B. M. Ali, P. Varahram, and A. J. Jara, "A Network Mobility Solution Based on 6LoWPAN Hospital Wireless Sensor Network (NEMO-HWSN)" in Seventh International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, Taichung Taiwan, 433-438, 2013. https://doi.org/10.1109/IMIS.2013.157A
[25] A. Dohr, R. Modre-Opsrian, M. Drobics, D. Hayn, G. Schreier, "The Internet of Things for ambient assisted living" in 7th International Conference on Information Technology: New Generations (ITNG 2010), Las Vegas USA, 804–809, 2010. https://doi.org/10.1109/ITNG.2010.104
[26] F. Goncalves, J. Macedo, M. J. Nicolau, A. Santos, "Security architecture for mobile e-health applications in medication control" in 21th International Conference on Software, Telecommunications and Computer Networks (SoftCOM 2013), Split - Primosten Croatia, 1–8, 2013, https://doi.org/10.1109/SoftCOM.2013.6671901
[27] L. Nie, T. Li, M. Akbari, J. Shen, T. Chua, "Wenzher: Comprehensive vertical search for healthcare domain" in ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Gold Coast Australia, 1245–1246, 2014. https://doi.org/10.1145/2600428.2611176
[28] H. Takeuchi, N. Kodama, "Validity of association rules extracted by healthcare-data-mining" in IEEE 36th Annu. Int. Conf. EMBC, Chicago USA, 4960–4963, 2014. https://doi.org/10.1109/EMBC.2014.6944737
[29] A. M. Ortiz, D. Hussein, S. Park, "The cluster between internet of things and social networks: Review and research challenges" IEEE Internet of Things Journal, 1.3, 206-215, 2014. https://doi.org/10.1109/JIOT.2014.2318835
[30] T. Y. Chung, I. Mashal, O. Alsaryrah, V. Huy, W. H. Kuo, D. P. Agrawal, "Social Web of Things: A Survey" in International Conference on Parallel and Distributed Systems (ICPADS), Seoul South Korea, 570-575, 2013. https://doi.org/10.1109/ICPADS.2013.102.
[31] H. Lee, J. Kwon, "Survey and Analysis of Information Sharing in Social IoT" in 8th International Conference on Disaster Recovery and Business Continuity (DRBC), Jeju South Korea, 15-18, 2015. https://doi.org/ 10.1109/DRBC.2015.13
[32] L. E. Holmquist, F. Mattern, B. Schiele, P. Alahutha, M. Beigl, H. Gallersen, "Smart-its friends: a technique for users to easily establish connections between smart artefacts" in ACM UbiComp'01, Atlanta USA, 2001. https://doi.org/10.1007/3-540-45427-6_10
[33] S. D. Bruyne, "Finding your way around the KNX Specifications" in KNX Technology Tutorial Workshop, Deggendorf Germany, 2004.
[34] B. A. Miller, T. Nixon, C. Tai, M. D Wood, "Home networking with Universal Plug and Play" IEEE Communications Magazine, vol. 39, no. 12, 104-109, 2001. https://doi.org/10.1109/35.968819
[35] https://www.knx.org/knx-en/software/ets/about/index.php
[36] http://www.bticino.com/solutions/home-automation/
[37] J. S. Lee, Y. W. Su, C. C. Shen, "A Comparative Study of Wireless Protocols: Bluetooth, UWB, ZigBee, and Wi-Fi" in IECON 2007 - 33rd Annual Conference of the IEEE Industrial Electronics Society, Taipei Taiwan, 46-51, 2007. https://doi.org/10.1109/IECON.2007.4460126
[38] B. Malinowsky, G. Neugschwandtner, and W. Kastner, "Calimero: Next Generation", in Konnex Scientific Conference, 2007.
[39] A. Bobek, H. Bohn, F. Golatowski, G. Kachel, A. Spreen, "Enabling workflow in UPnP networks" in INDIN'05 3rd IEEE International Conference Industrial Informatics, 166-171, 2005. https://doi.org/10.1109/INDIN.2005.1560370
[40] https://www.myopen-legrandgroup.com
[41] A. V. Aho, M. S. Lam, R. Sethi, J. D. Ullman, Compilers: Principles, Techniques, and Tools (2nd Edition), Addison-Wesley Longman Publishing Co. Inc., 2006.
[42] https://www.raspberrypi.org/products/raspberry-pi-3-model-b-plus/
[43] V. Miori, D. Russo "Home automation devices belong to the IoT world" ERCIM News, 101, Special theme: The Internet of Things and the Web of Things, ERCIM, 22 - 23, 2015.
[44] A. G. González, GNU General Public License v3: A Legal Analysis SCRIPT-ed, 155-163, 2006
[45] https://github.com/darioxrusso/smartSMILE

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 184-189 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis

Charles Frank[1], Asmail Habach[2], Raed Seetan[*,1], Abdullah Wahbeh[1]

[1] *Computer Science Department, Slippery Rock University, 16057, USA*

[2] *Mathematics Department, Slippery Rock University, 16057, US*

ARTICLE INFO

Article history:
Received: 23 February, 2018
Accepted: 11 March, 2018
Online: 27 March, 2018

Keywords:
Machine Learning
Smoker Status Prediction
Healthcare

ABSTRACT

Smoking has been proven to negatively affect health in a multitude of ways. As of 2009, smoking has been considered the leading cause of preventable morbidity and mortality in the United States, continuing to plague the country's overall health. This study aims to investigate the viability and effectiveness of some machine learning algorithms for predicting the smoking status of patients based on their blood tests and vital readings results. The analysis of this study is divided into two parts: In part 1, we use One-way ANOVA analysis with SAS tool to show the statistically significant difference in blood test readings between smokers and non-smokers. The results show that the difference in INR, which measures the effectiveness of anticoagulants, was significant in favor of non-smokers which further confirms the health risks associated with smoking. In part 2, we use five machine learning algorithms: Naïve Bayes, MLP, Logistic regression classifier, J48 and Decision Table to predict the smoking status of patients. To compare the effectiveness of these algorithms we use: Precision, Recall, F-measure and Accuracy measures. The results show that the Logistic algorithm outperformed the four other algorithms with Precision, Recall, F-Measure, and Accuracy of 83%, 83.4%, 83.2%, 83.44%, respectively.

1. Introduction

As of 2009, smoking has been considered the leading cause of preventable morbidity and mortality in the United States, continuing to plague the country's overall health [1]. Patients admitted to a hospital are often asked their smoking status upon admission, but a simple yes/no answer can be misleading. Patients who answer no can previously be smokers, or have recently quit smoking. The 'no' responses also do not consider their household member's smoking status, which can lead to continued exposure to secondhand smoke. Lastly, a 'no' response could still experience tobacco exposure through other forms, such as chewing tobacco. This study aims to use machine learning algorithms to predict a patient's smoking status based on medical data collected during their stay at a medical center. In the future, these predictive models may be useful for evaluating a patient's smoking status who is unable to speak.

Smoking has been proven to negatively affect your health in a multitude of ways. Smoking and secondhand smoke can magnify current harmful health conditions, and has been linked as the cause for others. Smoking and secondhand smoke often trigger asthma attacks for persons suffering from Asthma, and almost every case of Buerger's disease has been linked to some form of tobacco exposure. Various forms of cancer are caused by smoking, secondhand smoke, and other tobacco products [2]. In addition to being deemed the cause of certain cancers, most commonly known for causing lung and gum cancer, smoking and secondhand smoke also prevents the human body from fighting against cancer. Gum disease is often caused by chewing tobacco products, but continuing to smoke after gum damage can inhibit the body from repairing itself, including the gums. Smoking, secondhand smoke, and tobacco products are included in creating and preventing the recovery of the following additional diseases or health conditions: chronic obstructive pulmonary disease (COPD), diabetes, heart disease, stroke, HIV, mental health conditions such as depression and anxiety, pregnancy, and vision loss or blindness [3].

The objective of this study is to gain insights on smoking by exploring and studying patient's summary information after hospital admission. Using predictive machine learning models, a better understanding of tobacco's effect on a patient's health status can be obtained if the models produce valuable results. Using such models and existing patients' related information, understanding how vitals and patient data reflect the use of tobacco and smoking, could help medical professionals have better understanding of the

[*]Corresponding Author: Raed Seetan, Computer Science Department, Slippery Rock University, 16057, USA, Email: raed.seetan@sru.edu Tel: 724-738-2940

smokers population, which in turn can help better treat and handle patients with previous or current tobacco use more effectively.

Machine learning techniques are being applied to a growing number of domains including the healthcare industry. The fields of machine learning and statistics are closely related, but different in terms a number of terminologies, emphasis, and focus. In this work, machine learning is used to predict the smoking status of patients using several classification algorithms. Such algorithms include Multilayer Perceptron, Bayes Naïve, Logistic Regression, J48, and Decision Tree. The algorithms are used with the objective of predicting a patient's smoking status based on vitals. To determine if smoking has negative effects on vitals, One-way ANOVA analysis with SAS tool will be used repeatedly to determine whether different blood test readings from the patients are statistically different between smokers and non-smokers. The dataset used in this study was obtained from a community hospital in the Greater Pittsburgh Area [4]. The data set consists of 40,000 patients as well as 33 attributes.

The remainder of this paper is structured as follows: Section 2 discusses related work and the used dataset. Section 3 presents analytic methods and results. Section 4 discussed the results provides further recommendations; and Section 5 concludes the study.

2. Related Work

The I2b2 is a national center for Biomedical Computing based at Partners HealthCare System in Boston Massachusetts [5]. I2b2 announced an open smoking classification task using discharge summaries. Data was obtained from a hospital (covered outpatient, emergency room, inpatient domains). The smoking status of each discharge summary was evaluated based on a number of criteria. Every patient was classified as "smoker", "non-smoker", or "unknown". If a patient is a smoker, and temporal hints are presented, then smokers can be classified as "past smoker" or "current smoker." Summaries without temporal hints remained classified as "smoker".

Uzuner et. al. utilized the i2b2 NLP challenge smoking classification task to determine the smoking status of patients based on their discharge records [6]. Micro-average and macro-averaged precision, recall, and F-measure were metrics used to evaluate performance in the study. A total of 11 teams with 23 different submissions used a variety of predictive models to identify smoking status through the challenge with 12 submissions scoring F-measures above 0.84. Results showed that when a decision is made on the patent smoking status based on the explicitly stated information in medical discharge summaries, human annotators agreed with each other more than 80% of the time. In addition, the results showed that the discharge summaries express smoking status using a limited number of key textual features, and that many of the effective smoking status identifiers benefit from these features.

McCormick et. al., also utilized the i2b2 NLP challenge smoking classification task using several predictive models on patient's data to classify a patient's smoker status [7]. A classifier relying on semantic features from an unmodified version of MedLEE (a clinical NLP engine) was compared to another classifier which relied on lexical features. The classifiers were compared to the performance of rule based symbolic classifiers. The supervised classifier trained by MedLEE stacked up with the top performing classifier in the i2b2 NLP Challenge with micro-averaged precision of 0.90, recall of 0.89, and F-measure of 0.89.

Dumortier et. al. studied a number of machine learning approaches to use situational features associated urges to smoke during a quit attempt in order to accurately classify high-urge states. The authors used a number of classifiers including Bayes, discriminant analysis, and decision tree learning methods. Data was collected from over 300 participants. Sensitivity, specificity, accuracy and precision measures were used to evaluate the performance of the selected classifiers. Results showed that algorithms based on feature selection achieved high classification rates with only few features. The classification tree method (accuracy = 86%) outperformed the naive Bayes and discriminant analysis methods. Results also suggest that machine learning can be helpful for dealing with smoking cessation matters and to predict smoking urges [8].

3. Data Analysis and Results

The analysis is divided into two parts. In part 1, we use One-way ANOVA analysis with SAS tool to show the statistically significant difference in blood test readings between smokers and non-smokers. In part 2, we use five machine learning algorithms - Naïve Bayes, MLP, Logistic regression classifier, J48 and Decision Table - to predict the smoking status of patients. To compare the effectiveness of these algorithms we use four metrics, namely Precision, Recall, F-measure and Accuracy measures.

3.1. Statistical Analysis using ANOVA Test

In this work, One-way ANOVA analysis with SAS tool [9] was used repeatedly to determine whether different blood test readings from the patients are statistically different between smokers and non-smokers. So, our hypothesis are as follows.

Null Hypothesis (H0): There is no statistical difference in blood test readings between smokers and non-smokers.

Alternative Hypothesis (H1): There is a statistical difference in blood test readings between smokers and non-smokers.

The One-way ANOVA test, based on a 0.05 significance level, and the decision rule will be based on the p-value from the SAS outputs. If the p-value is less than 0.05, the null hypothesis is rejected and the alternative hypothesis is accepted. On the other hand, if the p-value is greater than 0.05, the null hypothesis is accepted. The analysis will be repeated for all blood tests, each of which is listed in Table 1 along with a brief description of its significance.

The results in Table 2 show that there is a significant statistical difference between smokers and non-smokers when it comes to three blood tests: INR, HB, and HCT. To investigate whether these differences were in favor of smokers or non-smokers, descriptive analysis was used (Figures 1, 2, and 3) to show the distribution of each blood test between smokers and non-smokers.

Figure 1 shows that non-smokers have higher values of INR than smokers. According to Mayo Clinic [10], an INR range of 2.0 to 3.0 is generally an effectiveness of anticoagulants. This shows

that non-smokers have a more effective therapeutic range than smokers.

Table 1: Lab value definitions

Blood Test	Significance
INR	Measures the effectiveness of the anticoagulants
Platelets	Involved in clotting
Glucose	Main source of energy and sugar
RBC	Red blood cells: carry oxygen and waste products
HB	Hemoglobin: Important enzyme in the RBCs
HCT	Hematocrit: measures the %RBC in the blood
RDW	Red blood cell distribution width

Table 2: Consolidated statistical results

Vital Reading	P-Value	Decision
INR	<0.0001	Reject the Null Hypothesis
Platelets	0.2935	Accept the Null Hypothesis
Glucose	0.1559	Accept the Null Hypothesis
RBC	0.0882	Accept the Null Hypothesis
HB	0.0005	Reject the Null Hypothesis
HCT	0.0022	Reject the Null Hypothesis
RDW	03509	Accept the Null Hypothesis

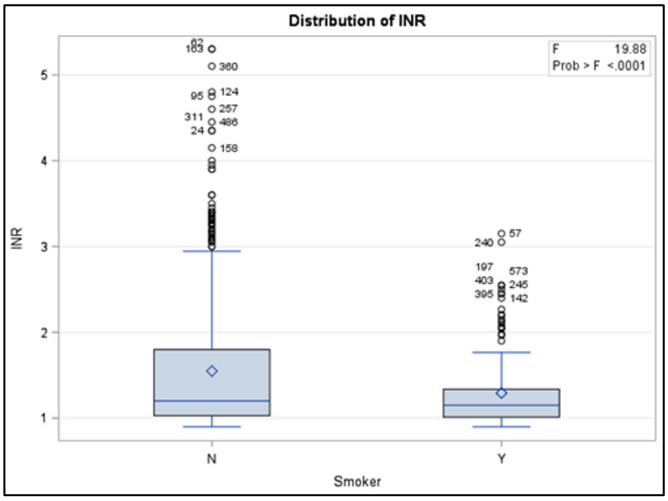


Figure 1: Distribution of INR blood test results between smokers and non-smokers

Figure 2 shows that non-smokers have lower values of HB than smokers. According to Mayo Clinic, an HB range between 12.0 and 17.5 is considered normal. This shows that although the readings of HB blood tests were statistically different between non-smokers and smokers, the difference was in general within the normal range.

Figure 3 shows that non-smokers have lower values of HCT than smokers. According to Mayo Clinic, an HB range between 37.0 and 52.0 is considered normal. This shows that although the readings of HCT blood tests were statistically different between non-smokers and smokers, the difference was in general within the normal range.

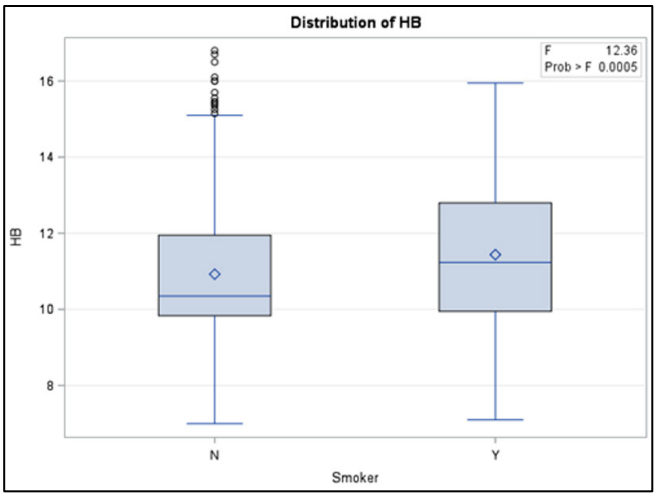


Figure 2: Distribution of HB blood test results between smokers and non-smokers

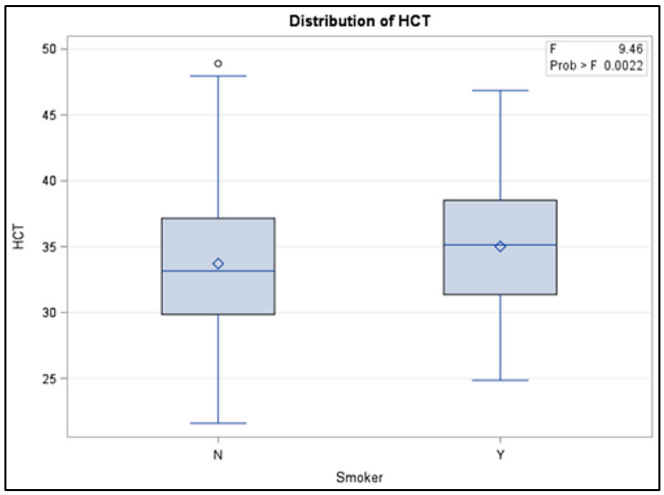


Figure 3: Distribution of HCT blood test results between smokers and non-smokers

3.2. Classification Analysis using Machine Learning

In this work, the Waikato Environment for Knowledge Analysis (Weka) (https://www.cs.waikato.ac.nz/ml/weka/) will be utilized to analyze the dataset [11]. The machine learning models utilized in this study include five classification algorithms, namely, Naïve Bayes, Multilayer Perceptron, Logistic, J48, and Decision Table.

3.2.1. Classifiers Description

Table 3 provides a summary about the classification algorithms characteristics and features. Naive Bayes is a popular versatile algorithm based on Bayes' Theorem, from the English mathematician Thomas Bayes. Bayes' Theorem provides the relationship between the probability of two events and the conditional probabilities of those events. The Naïve Bayes Classifier assumes that the presence of one feature of a class is not related to the presence or absence of another. Naïve Bayes classifier is a well-known algorithm because of its reputation for computational efficiency and overall predicative performance [12].

Table 3: Summary of classifiers characteristics and feature

Algorithm	Characteristics and Feature
Naïve Bayes	Computationally efficient, independence assumptions between the features, needs less training data, works with continuous and discrete data.
MLP	Many perceptrons organized into layers, ANN models are trained but not programmed, consist of three layers: input layer, hidden layer, and output layer.
Logistic	Multinomial logistic regression model with a ridge estimator
J48	Creates a binary tree, selects the most discriminatory features, and comprehensibility
Decision Table	Groups class instances based on rules, easy to understand, provides good performance

Multilayer Perceptron (MLP) is an Artificial Neural Network (ANN) model that maps sets of input data onto sets of suitable output data. ANN models are trained, not programmed. This means that the model takes a training set of data and applies what it has learned to a new set of data (the test data). The MLP ANN model is similar to a logistic regression classifier, with three layers: input layer, hidden layer, and output layer. The hidden layer exits to create space where the input data can be linearly separated. More hidden layers may be used for added benefit and performance, but MLP is used because of its overall performance [13].

The Logistic algorithm is a classifier for building and using multinomial logistic regression model with a ridge estimator to classify data. The version implemented using Weka states that it is slightly modified from the normal Logistic regression model, mainly to handle instance weights [14].

The J48 algorithm is a popular implementation of the C4.5 decision tree algorithm. Decision tree models are predictive machine learning models that determine the output value based on the attributes of input data. Each node of a decision tree signifies each attribute of the input data. The J48 model creates a decision tree that identifies the attribute of the training set that discriminates instances most clearly. Instances that have no ambiguity are terminated and assigned an obtained value, while other cases look for an attribute with the most information gain. When the decision tree is complete, and values are assigned to their respective attributes, target values of a new instance are predictively assigned [15].

Lastly, the Decision Table algorithm utilizes a simple decision table to classify data. Decision tables are best described to programmers as an if-then-else statement, and less complicated as a flow chart. A decision table groups class instances based on rules. These rules sort through instances and their attributes and classify each instance based on those rules. Decision tables are often easier to understand than other algorithm models while providing necessary performance [16].

Each model was run with a 66% split, using 66% of input as the training data and 34% as the test data. All algorithms are implemented through Weka after preprocessing, Table 4.

Table 4: Weka Schema

Algorithm	Weka Schema Attribute
Naïve Bayes	weka.classifiers.bayes.NaiveBayes
MLP	weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 5 -E 20 -H a
Logistic	weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4
J48	weka.classifiers.trees.J48 -C 0.25 -M 2
Decision Table	weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

3.2.2. Preprocessing

A sample of the large dataset was used for analysis due to available resources. A few samples were created prior to this study for previous work. The sample contains 534 total patients, with 311 non-smokers and 87 smokers. This remains relatively consistent with the overall ratio of smokers to non-smokers in the full dataset. The smoker attribute contained 136 missing values, accounting for 25% of the patients in the sample. To account for missing values, we utilized a Weka filter called "ReplaceMissingValues". This filter replaces missing values through the selected attribute with modes and means of the values in the training set.

After addressing missing values for the class attribute, oversampling was applied to add additional data for analysis. Oversampling was added to try and alter the ratio of smokers to non-smokers closer to the original dataset. The SMOTE algorithm was used through Weka and applied three times, bringing the total instances to about 1000 patients. All preprocessing is done using Weka as listed in Table 5.

Table 5: Weka Filters

Algorithm	Type	Weka Attribute
SMOTE	Filter	weka.filters.supervised.instance.SMOTE
ReplaceMissing Value	Filter	weka.fliters.unsupervised.attribute.-ReplaceMissingValue

3.2.3. Means of Analysis

To evaluate the performance of the machine learning model's, four different measures are used, namely: Precision, Recall, F-measure, and Accuracy, shown in equations 1-4. Precision shows the percent of positive marked instances that truly are positive. Recall is the percentage of positive instances that are correctly identified. Recall is also referred to as sensitivity. F-measure or F-score is a measure of accuracy, that considers the harmonic mean of precision and recall. Accuracy is simply the amount of correctly classified instances from an algorithm.

$$Precision = \frac{TP}{TP + FN} \quad (1)$$

$$Recall = \frac{TN}{TN + FP} \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (3)$$

$$\mathrm{F} - measure = 2 . \frac{Precision . Recall}{Precision + Recall} \quad (4)$$

Figures 4 shows the performance of the five algorithms using the Precision measure. Results show that the J48 and Logistic achieved the highest precision with 83%, followed by MLP, Decision Tree, and Naïve Bayes with Precision values of 81%, 80.5%, and 77.8% respectively.

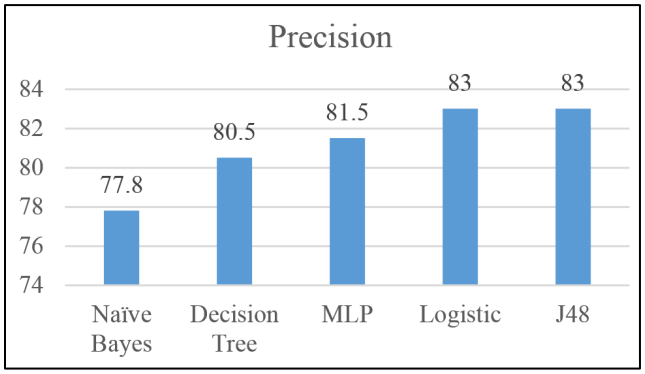


Figure 4: Precision for Naïve Bayes, MLP, Logistic, J48, and Decision Tree results

Figures 5 shows the performance of the five algorithms using the Recall measure. Results show that the J48 achieved the highest Recall with 83.4%, followed by Logistic, MLP, Decision Tree, and Naïve Bayes with Recall values of 83.1%, 81.8%, 81.1% and 77.8% respectively.

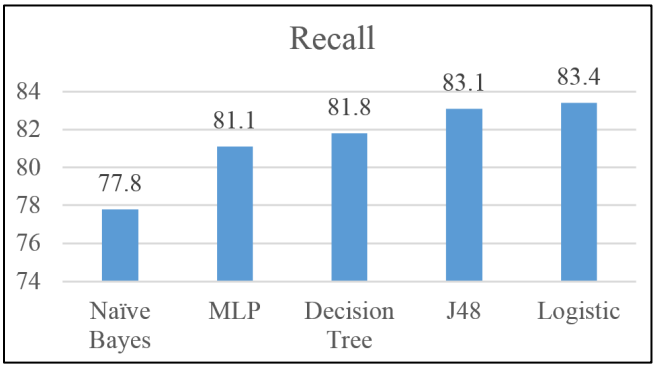


Figure 5: Recall for Naïve Bayes, MLP, Logistic, J48, and Decision Tree results

Figures 6 shows the performance of the five algorithms using the F-Measure. Results show that the Logistic achieved the highest F-Measure with 83.2%, followed by J48, MLP, Decision Tree, and Naïve Bayes with Recall values of 83.1%, 81.3%, 81% and 78.8% respectively.

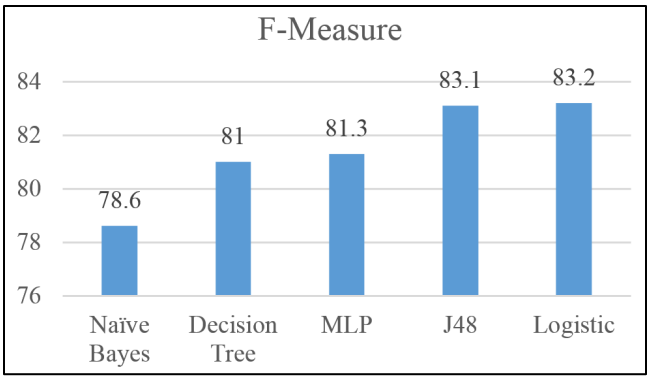
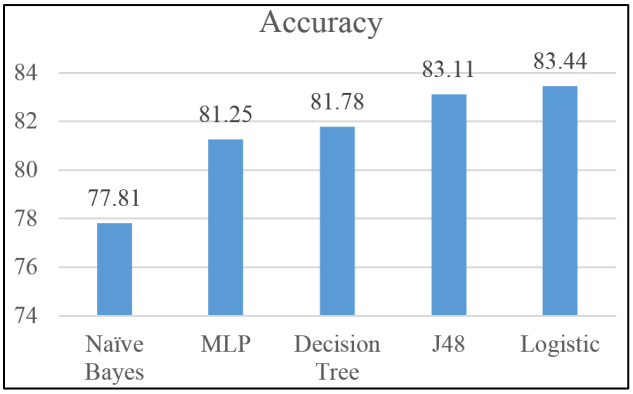


Figure 6: F-measure for Naïve Bayes, MLP, Logistic, J48, and Decision Tree results



Overall, the results show an indication that the five algorithms are relatively reliable when it comes to predicting the smoking status of patients. Logistic algorithm outperformed the four other algorithms with Precision (83%), Recall (83.4%), F-Measure 83.2%, and Accuracy (83.44%).

The study addressed the potential of machine learning algorithms to predict the status of smoking among a smoker population. Results showed the potential of such algorithms to predict the smoking states with accuracy level of 83.44%. However, this study has few limitations. There are several items that could be addressed to further this study and improve outcomes, beginning with data preprocessing. Several other methods are available to handle missing values in the dataset. In this study, the ReplaceMissingValues filter was applied in Weka to handle missing values. The Weka Filter replaces null values with means and modes from the training set. Using a method such as replacing null values with moving averages could produce results that are more realistic to the actual smoking status of patients. Other methods of handling missing values could also be explored.

Another improvement could be to increase the size of the sample dataset. In this study, a sample of 534 patients was used and the SMOTE model was applied in preprocessing. Using a larger sample set could also bring the results closer to what would be expected when applying these tests to larger sets of patients. In most cases, there will be more than 534 patient entries to analyze.

So, learning the results of these tests on larger real-world sets could further prove the value of these tests.

Lastly, other models may outperform those tested in this study. While five algorithms were tested, and showed wholesome results, others may provide better marks. Clustering models may be a point of interest as those tested in this study are all classifier models. There is an abundant amount of classifier models out there and their results are worth testing.

5. Conclusion

This study showed that five machine learning models can be used reliably to determine the smoking status of patients given blood tests and vital readings attributes. These algorithms are Naïve Bayes, MLP, Logistic, J46 and Decision Tree. Logistic algorithm outperformed the other four algorithms with precision, recall, F-Measure, and accuracy of 83%, 83.4%, 83.2%, 83.44%, respectively.

Using One-way ANOVA analysis with SAS tool, the study also confirmed that there is a significant statistical difference between smokers and non-smokers when it comes to three blood tests: INR, HB, and HCT. The difference was within the normal range with HB and HCT, but it was in favor of non-smokers with INR which measures the effectiveness of anticoagulants. In the future, the models could be implemented in hospital systems to identify patients who do not specify smoking status. Also, the findings from SAS confirms the negative health effects of being a smoker.

References

[1] Centers for Disease Control and Prevention: https://www.cdc.gov/tobacco/campaign/tips/diseases/?gclid=CjwKEAjw5_vHBRCBtt2NqqCDjiESJABD5rCJcbOfOo7pywRlcabSxkzh0VIifcvYI05u-hQ9SsI9RRoCDZfw_wcB

[2] The Mayo Clinic (2017). Retrieved May 01, 2017, from http://www.mayoclinic.org/

[3] S., Dube, A., McClave, C., James, R., Caraballo, , R., Kaufmann, & T., Pechacek, (2010, September 10). Vital Signs: Current Cigarette Smoking Among Adults Aged =18 Years United States, 2009. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5935a3.htm

[4] M., Keyes, C., Frank, A., Habach and R. Seetan. Artificial Neural Network Predictability: Patients' Susceptibility to Hospital Acquired Venous Thromboembolism. The 32th Annual Conference of the Pennsylvania Association of Computer and Information Science Educators (PACISE), At Edinboro University of Pennsylvania. March 31st and April 1st, 2017.

[5] Partners Healthcare. (2017). Informatics for Integrating Biology & the Bedside. Retrieved May 23, 2017, from i2b2: https://www.i2b2.org/NLP/DataSets/Main.php

[6] O., Uzuner, I., Goldstein, Y., Luo, & I., Kohane, (2008). Identifying Patient Smoking Status from Medical Discharge Records. J Am Med Inform Assoc, 15(1), 14-24. doi:10.1197/jamia.m2408

[7] P., McCormick, N., Elhadad & P., Stetson (2008). Use of Semantic Features to Classify Patient Smoking Status. Retrieved March 5, 2017, from Columbia.edu: http://people.dbmi.columbia.edu/noemie/papers/amia08_patrick.pdf

[8] A., Dumortier, E., Beckjord, S., Shiffman,, & E., Sejdić (2016). Classifying smoking urges via machine learning. Computer methods and programs in biomedicine, 137, 203-213.

[9] SAS Institute Inc., SAS 9.4 Help and Documentation, Cary, NC: SAS Institute Inc., 2017.

[10] The Mayo Clinic (2017). Retrieved May 01, 2017, from http://www.mayoclinic.org/

[11] E., Frank, M., Hall, and I., Witten (2016). The Weka Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[12] E., Frank, & R., Bouckaert, (2006, September). Naive bayes for text classification with unbalanced classes. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 503-510). Springer, Berlin, Heidelberg.

[13] W., Gardner, and S. Dorling. "Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences." Atmos. Environ. 32, no. 14-15 (1998): 2627-2636.

[14] W., Hosmer, S., Lemeshow, and R., Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[15] D., Dietrich, B., Heller, and B., Yang. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. (2015). Wiley.

[16] H., Lu, & H., Liu, (2000). Decision tables: Scalable classification exploring RDBMS capabilities. In Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'00 (p. 373).

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 186-199 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

NemoMap: Improved Motif-centric Network Motif Discovery Algorithm

Tien Huynh[*,1], Somadina Mbadiwe[2], Wooyoung Kim[1]

[1]*Computing and Software Systems, School of STEM, University of Washington Bothell, Bothell, WA 98011, USA*

[2]*Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA*

ARTICLEINFO

Article history:
Received: 16 August, 2018
Accepted: 17 September, 2018
Online: 29 September, 2018

Keywords:
Network Motif
Motif-centric
Grochow-Kellis
MODA

ABSTRACT

Network motif analysis has several applications in many different fields such as biological study and social network modeling, yet motif detection tools are still limited by the intensive computation. Currently, there are two categories for network motif detection method: network-centric and motif-centric approach. While most network-centric algorithms excel in enumerating all potential motifs of a given size, the runtime is infeasible for larger size of motifs. Researchers who are interested in larger motifs and have established a set of potential motif patterns could utilize motif-centric tools to check whether such patterns are truly network motifs by mapping them to the target network and counting their frequency. In the paper, we present NemoMap (Network Motif Mapping algorithm) which is an improvement of the motif-centric algorithm, GK (by Grochow and Kellis) and MODA (Motif Detection Algorithm). Experimental results on three different protein-protein interaction networks show that NemoMap is more efficient in mapping complex motif patterns, while GK and MODA is much faster in analyzing simpler patterns with fewer edges. We also compare the performance of NemoMap and ParaMODA (introduced previously to improve MODA), and the result shows that NemoMap yields better runtime due to the implementation of Grochow-Kellis' symmetry-breaking technique and the better node selection process.

1 Introduction

Advancements in modern computer processing power and storage capability have enabled the collection of large-scale networks such as social networks, computer networks, and biological networks. These networks hold thousands or even millions of nodes (vertices), and they could be modelled using the graph structure which allows researchers from different fields to study under a common framework. One of the biggest challenges in network analysis is to determine network motifs, defined as the subgraph patterns that occur more frequently in the original network than in similar randomized networks. As an extension of work originally presented in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017 [1], we provide an improved network motif detection algorithm in this paper.

The motivation for network motif study comes from the desire to gain more insight on the structural foundation of complex networks because those substructures might be the key component of a specific functionality. Such knowledge could be used to classify networks into "superfamilies" [2], or to select appropriate network model to study a real network [3]. In biological networks, network motif analysis has been used for discovering basic functional foundation in Transcriptional Regulation Networks (TRN) [4, 5], predicting interaction in Protein-Protein Interaction (PPI) networks [6], and studying breast-cancer related genes [7].

Although network motif analysis has significant ap-

[*]Corresponding author, Computing and Software Systems, School of STEM, University of Washington Bothell, WA 98011, USA, huynhti@uw.edu

https://dx.doi.org/10.25046/aj030523

plications in the real world, in practice it is infeasible to find motif of larger size because the process involves a nondeterministic polynomial time (NP) isomorphic testing and statistical testing. Two graphs which have the same number of nodes connecting in the same way are considered isomorphic, and most network motif analysis includes a huge number of isomorphic testing to classify motif and to avoid double counting. Additionally, statistical testing is used to determine the uniqueness of motifs in a number of similar randomly generated networks and the same process of isomorphic testing runs for each newly created network (or ensemble of networks), so time complexity increases dramatically. Typically, more than $1,000$ networks are generated, and the frequency of each motif is statistically compared using P-value or Z-score. A subgraph pattern with P-value of less than 0.01 or a Z-score of higher than 2 is identified as a network motif [8]. Such operations are very time consuming and usually are the main bottleneck of network motif analysis.

There have been several motif detection algorithms developed over the years, and they generally fall into two categories: network-centric and motif-centric [9]. The network-centric approach would search all possible enumerations of subgraphs of a given size in the target network, while the motif-centric method would predetermine one or more query graph patterns then count the occurrence of each query graph in the original network. Both approaches utilize statistical testing to check the significance of each potential motif.

Here, we propose NemoMap (Network Motif Mapping algorithm) which is an extension of ParaMODA in [1]. NemoMap improves the performance of ParaMODA which is an improvement of the motif-centric algorithms developed by Grochow-Kellis (GK) [10]. NemoMap is implemented in C++ so that it can be included in NemoLib [11], which is currently developed with Java, C++ and Python and includes network-centric method only. With the addition of NemoMap in NemoLib, we expect to have both choices available in a single library. In this paper, we also compare the performance of NemoMap against GK, as well as ParaMODA.

The rest of the paper is organized as the following: Section 2 describes background information of network motif; Section 3 discusses NemoMap algorithm and its implementation; Section 4 reports experimental results and performance comparison, followed by conclusion and future works in Section 5.

2 Network Motif

2.1 Network Motif Detection

Network motifs were introduced in 2002 by Milo et al. as "patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks" [4]. Suppose we have a network G with $|G|$ vertices, and a number k such that $3 \leq k << |G|$. A connected subgraph M of size k which appears more frequently in G than a predetermined threshold value is a network motif. Typically, we use the statistical P-value or Z-score of the frequency of M in thousands of randomly generated similar networks as the threshold value.

$$\text{P-value}(M) = \frac{1}{N}\sum_{n=1}^{N} c(n), \tag{1}$$

$$\text{where } c(n) = \begin{cases} 1, \text{if } f_R(M) \geq f_G(M). \\ 0, \text{otherwise} \end{cases}$$

$$\text{Z-score}(M) = \frac{f_G(M) - \mu(f_R(M))}{\sigma(f_R(M))} \tag{2}$$

Here, N is the number of random graphs, and $f_G(M)$ is the frequency of M in the target network while $f_R(M)$is the frequency of M in the random network. $\mu(f_R(M))$ is the mean of frequencies of M in the random networks and $\sigma(f_R(M))$ is the standard deviation of frequencies of M in the random networks.

Generally, a subgraph with P-value < 0.01 or Z-score > 2.0 is considered a network motif after analyzing in more than $1,000$ random networks [8].

2.2 Isomorphic Testing

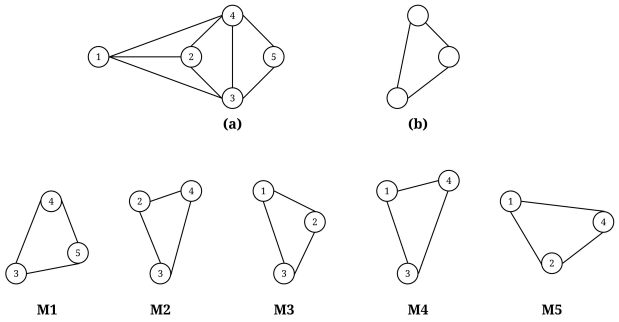


Figure 1: (a) A sample undirected graph; (b) A subgraph pattern; (c) M1-M5 are all isomorphic to (b)

Two graphs are isomorphic if they have the same number of nodes and there is "a one-to-one mapping between their nodes such that each edge in one graph can be mapped to an edge in the other graph" [12]. Figure 1 shows isomorphic subgraph patterns as an example. Since isomorphic testing is known to be an NP-complete problem, a common practice for categorizing graphs is to use a program called *Nauty* [13] to canonically label subgraph patterns; if two graphs have the same canonical label, they are isomorphic. One drawback of using Nauty is its dependency on the Linux operating system.

Another technique for analyzing isomorphism is comparing certain invariants of the graphs such as the degree of each node or the degrees of each node's neighbors [10]. Those properties are inherent to the structure of the graph, so two graphs are isomorphic if

they have the same properties. Here, we will utilize the later approach because we want the implementation without Nauty, to be operating system (OS) independent.

2.3 Network-Centric Approach

Many researchers have developed algorithms and software implementations of network motif detection focusing on the network-centric approach which searches all possible enumerations of subgraphs of a given size in the original target network. Some of the most distinguished applications include MAVisto [14], MFinder [15], FANMOD [16, 17], Kavosh [18], NemoFinder [19], and NetMODE [20], most of which involve the typical process of enumerating different k-size subgraphs in the original network. Unfortunately, even with modern computer architecture such solutions are still limited in the size of detectable motifs. Network-centric algorithms using exhaustive exact enumeration of k-size subgraphs in the network can often find only motifs with up to 6 nodes [10] due to the time-consuming process of scanning for every nodes and edges of the network.

One method to substantially reduce such processing time is sampling, employed by applications like MFinder and FANMOD. MFinder uses an edge-sampling strategy introduced by Kashtan et al. [15] which randomly selects an edge in the target network then picks a k-size subgraph around that edge to compare. However, edge-sampling technique seems to be biased toward subgraphs that have more edges resulting in higher count for such subgraphs [12]. Additionally, there is also the node-sampling strategy used by FANMOD which assigns a probability to the node to determine the chance that such node would be explore further [17]. The later sampling technique proves to be more efficient in reducing runtime, yet it could only increase FANMOD maximum discoverable motif size to 8 nodes.

2.4 Motif-Centric Approach

Due to the limitation of network-centric algorithms, Grochow and Kellis (GK) introduced a different approach for larger motif detection by first determining a query list of subgraph patterns of size k, then finding the frequency of each pattern in the target network by mapping the query graphs to all possible location in the network [10].

This strategy is described as "motif-centric", and the application can find motifs of up to 15 nodes. Moreover, the algorithm employs an efficient enumeration process called symmetry-breaking technique to significantly reduce isomorphic testing which is one of the most intensive tasks in network motif detection. The symmetry-breaking condition ensures that each mapping of the query graph to the target network is unique so that unnecessary isomorphic testing is avoided. However, one disadvantage is that it requires a set of query subgraphs to start with. Therefore, if the researcher has no method to determine an appropriately concise k-size query list, run-time can suffer heavily with increasing subgraph size since full enumeration of all possible variations of k-size subgraphs is needed. Table 1 shows the exponential growth of pattern variations of k-size subgraphs up to 10 nodes, and in the worst-case scenario the application would have to iterate over 341, 247, 400, 399, 400, 000, 000 of 10-node patterns many of which might not even exist in the target graph [9].

Table 1: Number of non-isomorphic subgraphs for undirected and directed graphs with up to 10 vertices. Courtesy of [8]

Vertices	Number of non-isomorphic subgraphs	
	Undirected	Directed
1	1	1
2	1	2
3	2	13
4	6	199
5	21	9,364
6	112	1,530,843
7	853	880,471,142
8	11,117	1,792,473,955,306
9	261,080	13,026,161,682,466,200
10	11,716,571	341,247,400,399,400,000,000

Heavily inspired by Grochow-Kellis' (GK) algorithm, MODA (Motif Discovery Algorithm) [21] improves upon the former by the introduction of a query graph hierarchy expansion treeas illustrated in Figure 2. MODA chooses a subgroup of the query graphs which might appear more frequently in the target network to start mapping with as the top level of the tree; then it adds one edge to each graph of the previous level while ensuring that no duplicate is present to build the subsequent level of the tree. Because each level is only extended by one edge, the mapping information of the previous level can be saved and reused to extend mapping by one node, so computation time would be reduced substantially. It also uses a node-sampling method during the edge-expansion based on the linearly proportional distribution of nodes and edges [9].

3 NemoMap

We propose NemoMap (Network Motif Mapping algorithm) as an extension of ParaMODA [1] which is an improvement of the motif-centric algorithm introduced by Grochow-Kellis (GK). We call it NemoMap since it is an improvement of the mapping function introduced by GK and we will include it in the NemoLib library [11]. Since this is a direct improvement of GK, we will firstly explain in detail the main components of GK, then discuss our alteration in NemoMap, and

lastly compare the difference in implementation of NemoMap and ParaMODA.

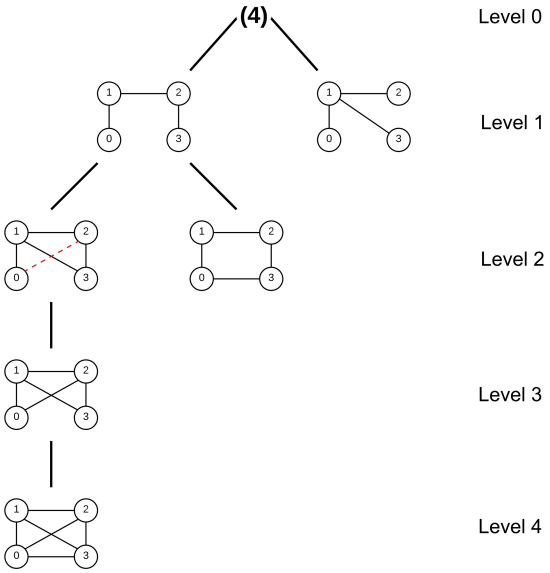


Figure 2: Expansion tree of size-4 subgraph, courtesy of [21]

3.1 Grochow-Kellis (GK)

We will discuss the mapping function of GK which focuses on the mapping of a query graph to the target network to find the frequency of such query graph's appearance in the network because NemoMap and ParaMODA improve upon such functionality.

Suppose we have an input network G and a query graph H, and the goal is to find and count all the possible mapping f from H to G which G can support. GK mapping algorithm involves three functions as shown in Algorithm 1, 2, and 3.

As discussed in Section 2.4, GK utilizes comparison of two invariants between graphs to test for isomorphism: the degree of each node and the degrees of each node's neighbors. It starts with an arbitrarily partial map of one node h of H to one node g of G, then extends that map by a single node for every recursive call of IsomorphicExtensions (Algorithm 2) until all nodes of H are mapped to appropriate nodes of G. Additionally, IsomorphicExtensions guarantees that any newly mapped node is connected to the already-mapped nodes, so the returned mapping must be an isomorphism [10].

Although the symmetry-breaking condition is optional, it significantly improves runtime and memory usage. Figure 3 shows symmetry-breaking example. Without symmetry-breaking, FindSubgraphInstances (Algorithm 1) would have to maintain a list of all mappings in memory to screen out duplicate mappings resulting in low efficiency as the program has to constantly write and read data from memory or disk. On the other hand, symmetry-breaking ensures that each mapping is unique, so there is no need to maintain the mapping list. Such feature substantially reduces memory usage and runtime for larger network analysis or bigger size query graphs [10] as illustrated in Table 2.

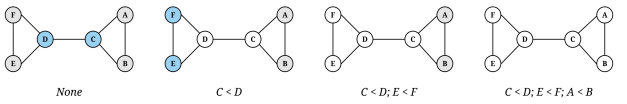


Figure 3: Symmetry-breaking example: Finding conditions that will break all the symmetries of a 6-node graph. White nodes are fixed by any automorphism preserving the indicated conditions, and other nodes are shaded according to their equivalence class under the automorphisms which preserve the indicated conditions. (Courtesy of [10])

3.2 ParaMODA

Compared to GK algorithm, ParaMODA follows closely with a couple of alterations. In FindSubgraphInstances in Algorithm 1, instead of going through all nodes h of H, we only arbitrarily choose one node h to traverse from. The chosen h node serves as the fixed starting point for all traversals of the query graph, and since the nodes of G are only traversed once with the chosen node h, the visited nodes g of G are not removed from the network [1]. There are no changes to IsomorphicExtensions in Algorithm 2.

ParaMODA improves time complexity by eliminating the inner iteration (loop) in FindSubgraphInstances at the expense of not shrinking the network by keeping visited nodes. The performance comparison will be discussed in Section 4. The logic behind the removal of the inner iteration comes from the fact that if the query graph is connected and the network is unchanged, any node of the network could be used as the starting point for mapping because the set of edges remain unchanged [1].

Another advantage of ParaMODA is the possibility of parallelization. Since each mapping from H to G is unique and each starting mapping node of the network is independent, we can partition the vertices easily across processors and have the result aggregated. Because the nodes are ordered based on their degrees in descending order, we can achieve better load distribution across processes by using strategies that factor in the uneven nature of that distribution [1].

3.3 NemoMap

As an extension to ParaMODA [1], NemoMap maintains the core structure with two differences: the clarification of node h choosing process, and the implementation of symmetry-breaking technique. We implement NemoMap in C++ so that it can be included in the NemoLib library [11].

Instead of arbitrarily picking a node h of H to start the mapping process as discussed in Section 3.2, NemoMap chooses the most constrained node h to traverse from. This selection process starts with picking the node with the highest degree, and if there are more than one nodes with the same highest degree we will

Algorithm 1: FindSubgraphInstances (G,H), by courtesy of [10]

Find all instances of query graph H in network G
Input : Graph G and a query graph H
Output: A set of all instances of H in G
Start with an empty set of instances
Find $Aut(H)$. Let H_E be the equivalence representatives of H
Find symmetry-breaking conditions C of H given H_E and $Aut(H)$
Order the nodes of G by increasing degree and then by increasing neighbor degree sequence
for *each node g of G* **do**
  for *each node h of H such that g can support h* **do**
    Let f be the partial map associating $f(h) = g$
    Find all isomorphic extensions of f [up to symmetry] (i.e. call `IsomorphicExtensions` (f,H,G, $C[h]$))
    Add the images of these maps to the set of all instances
  Remove g from G
Return the set of all instances.

Table 2: The number of subgraphs encountered by NemoMap with and without symmetry-breaking (including multiple encounters for the version without symmetry-breaking). The improvement factor is exactly the average number of automorphisms of subgraphs of the associated size. (Courtesy of [10])

	Undirected PPI Network			Directed Regulatory Network		
Nodes	Total Subgraphs Searched	With Symmetry-Breaking	Improvement	Total Subgraphs Searched	With Symmetry-Breaking	Improvement
2	3.7×10^4	1.1×10^4	×3.13	2.6×10^4	1.3×10^4	×2.02
3	4.0×10^5	7.0×10^4	×5.77	9.7×10^5	1.8×10^5	×5.41
4	4.4×10^6	4.1×10^5	×10.9	4.4×10^7	2.5×10^6	×18.0
5	5.1×10^7	2.3×10^6	×22.2	2.3×10^9	3.2×10^7	×73.3
6	5.7×10^8	1.2×10^7	×46.3	1.3×10^{11}	4.0×10^8	×334
7	6.4×10^9	6.6×10^7	×96.2	-	-	-

select among them the node with the highest sum of neighbors degrees. If there is still a tie after the second selection round, we will arbitrarily choose one node among the last filtered group to be node h. Such selection process helps reduce the number of recursive isomorphic testing as the constrained node h has a lower chance of being mapped to nodes in G, resulting in a faster runtime. Algorithm 4 shows the altered process.

Without symmetry-breaking, ParaMODA suffers from memory overhead issue when analyzing query with a large number of mappings because the program has to maintain a list of mappings to avoid duplication. As a result, the program frequently runs out of memory and crashes if the target network is large. Furthermore, it also has slower runtime as redundant recursive calls for isomorphic testing are completed. NemoMap solves this problem by including the GK symmetry-breaking condition with a little tweak. Because only one node h of H is chosen, SymmetryConditions in Algorithm 3 does not need to find the symmetry-breaking conditions for all nodes of H; only the conditions of the selected node h are needed as illustrated in Algorithm 5.

NemoMap implementation also uses the standard vector data structure instead of the user-defined class Mappings as in ParaMODA program. The change is due to performance optimization, and from the fact that all query graph mappings always start from one fixed node h with the same extension path. Therefore, from the standpoint of a key-value pair of the Mappings class with the keys representing mapped nodes h of query graph H, the key's sequence order will always be the same, so there is no need to maintain such data; we can save it at the beginning for later use if necessary. Instead, we use a vector to store the values, i.e. the corresponding mapping nodes g of network G, in the correct order that the keys are in.

The input file format for the query graph and the target network of NemoMap remains the same with each line contain two integers separated by a white space or a tab. Each line of the input file represents an edge of the graph, and the two integers are the ID number of the nodes which form such edge. Lines start with "#" are comments and will be ignored [1]. Additionally, the program disallows self-edge, i.e. edge from a node pointing to itself, as well as duplicating edge.

As discussed earlier, the program generates a list of mappings of the query graph H in all possible lo-

Algorithm 2: IsomorphicExtensions(f,H,G, $C[h]$), by courtesy of [10]

Find all isomorphic extensions of partial map $f : HG$ **[satisfying** $C(h)$**]**
Input : Partial map $f : H \rightarrow G$, a graph G, query graph H, and a symmetry-breaking condition $C(h)$
Output: All isomorphic extensions of f
Start with an empty list of isomorphisms
Let D be the domain of f
if $D = H$ **then**
  Return a list consisting solely of f. (Or write to disk)
Let m be the most constrained neighbor of any $d \in D$ (constrained by degree, neighbors mapped, etc.)
for *each neighbor n of* $f(D)$ **do**
  if *there is a neighbor* $d \in D$ *of* m *such that* n *is not neighbors with* $f(d)$*, or if there is a non-neighbor* $d \in D$ *of* m *such that* n *is neighbors with* $f(d)$*, or if assigning* $f(m) = n$ *would violate a symmetry-breaking condition in* $C(h)$*]* **then**
    Continue with the next n
  else
    Let $f' = f$ on D, and $f'(m) = n$
  Find all isomorphic extensions of f'
  Append these maps to the list of isomorphisms
Return the list of isomorphisms.

Algorithm 3: SymmetryConditions($H_E, Aut(H)$), by courtesy of [10]

Find symmetry-breaking conditions for H **given** H_E**,** $Aut(H)$
Input : Nodes of H grouped into equivalence classes H_E, automorphisms of H $Aut(H)$
Output: A map from equivalence representatives to sets of symmetry-breaking conditions
Let M be an empty map from equivalence representatives to sets of conditions
for *each* $n \in H_E$ **do**
  Let C be an empty set of conditions
  $n' = n$
  $A = Aut(H)$
  do
    Add "Label(n') <Min{Label(m)|$m \sim An'$ and $m \neq n'$}" to C
    $A = \{f \in A | f(n') = n'\}$
    Find the largest A-equivalence class E
    Pick $n' \in E$ arbitrarily
  while $|A| \neq 1$
  Let $M(n) = C$
Return M

cations of the network G. The mappings can be saved to disk if necessary, however the current NemoMAp implementation simply discards them and only count the number of mappings found.

3.4 Summary of Changes

The changes in NemoMap vs. GK algorithm are picking one fixed node h in terms of mapping and finding symmetry-breaking condition of only that node. NemoMap is implemented in C++ and includes the symmetry-breaking technique, while ParaMODA is written in C# and does not have symmetry-breaking; NemoMap also have an improved node h selection process. Table 3 summarizes the changes.

4 Experiments

4.1 Setup

Motif-centric approach is not primarily designed for exhaustive search of network motif, so we will focus on comparing runtime performance between NemoMap and GK on predetermined query graph mapping. The experiment will include two parts: performance comparison between NemoMap and ParaMODA, and between NemoMap and GK. All testing will be conducted on three protein-protein interaction (PPI) networks obtained from the Database of Interacting Proteins (DIP) [22]:

- *Homo sapiens* (Hsapi): 1,715 nodes with 1,873 edges (low edge density)
- *Escherichia coli* (E. coli): 1,223 nodes with 1,654 edges (average edge density)
- *Saccharomyces cerevisiae* (S. cerevisiae): 2,164 nodes with 4,303 edges (high edge density)

As mentioned in [1], the performance difference between ParaMODA and GK algorithm varies widely depending on the query graph used. In this paper, we attempt to elaborate which query graph categories

Algorithm 4: FindSubgraphInstances_NemoMap(G, H)

Find all instances of query graph H in network G
Input : Network G and a set of query graph H
Output: A set of all instances of H in G
Start with an empty set of instances
Pick the most constrained node h from H (constrained by degree, sum of neighbors' degrees)
Find $Aut(H)$
Find symmetry-breaking conditions C of h given $Aut(H)$
Order the nodes of G by increasing degree and then by increasing neighbor degree sequence
for *each node g of G such that g can support h* **do**
  Let f be the partial map associating $f(h) = g$
  Find all isomorphic extensions of f[up to symmetry] (i.e. call `IsomorphicExtensions` (f,H,G, $C[h]$))
  Add the images of these maps to the set of all instances
Return the set of all instances.

Algorithm 5: SymmetryConditions_NemoMap($H_E, Aut(H)$)

Find symmetry-breaking conditions for H given H_E, $Aut(H)$
Input : Automorphisms of $H, Aut(H)$
Output: A set of symmetry-breaking conditions for node h
Let C be an empty set of conditions
$n' = n$
$A = Aut(H)$
do
  Add "Label(n') <Min{Label(m)|$m \sim An'$ and $m \neq n'$}" to C
  $A = \{f \in A | f(n') = n'\}$
  Find the largest A-equivalence class E
  Pick $n' \in E$ arbitrarily
while $|A| \neq 1$
Let $M(n) = C$ Return M

are more advantageous to each algorithm. Therefore, we break down the query graphs into three pattern groups:

- Group Simple: Low edge-to-node count
- Group Average: Average edge-to-node count
- Group Complex: High edge-to-node count

We choose to use the ratio of edge count to node count as a measure for a graph complexity with higher ratio equating to higher complexity. Although there are other factors in a graphs complexity such as out-degree distribution, we think using the edge-to-node ratio would making the experiment more manageable. Each query graph group consists of four query graphs with four to seven nodes. All tests were conducted on a desktop computer with the following specifications:

- CPU: Intel Core i5-8400 (6-core) 2.80Ghz, Turbo 4.00Ghz, 9MB cache
- RAM: 16GB DDR4 2133Mhz
- Hard Drive: 240GB SSD
- GPU: Intel UHD Graphics 630
- OS: Windows 10 Home Edition

Table 3: Summary of changes of NemoMap compared to ParaMODA and GK

	NemoMap	**ParaMODA**	**GK**
Language	C++	C#	Java
Mapping	Pick the **most constrained node** h of query graph H (constrained by degree, sum of neighbors' degrees), then start all mapping from the h, and do not remove mapped node g	Pick an **arbitrary node** h of query graph H, then start all mapping from the h, and do not remove mapped node g	Iterate over **all nodes of** H (or **all representative node of** H if symmetry-breaking is used), and remove mapped node g after every iteration
Symmetry-breaking	Find symmetry-breaking condition for **only one fixed node** h	Symmetry-breaking condition is **not included**	Find symmetry-breaking condition for **all representative nodes of** H

- IDE: Visual Studio 2017 Community

The source code of Grochow-Kellis' algorithm in Java was provided by the author [10], and all runtimes reported for GK are measured using this implementation. The source code was compiled with the IntelliJ IDEA 2018.5 Community IDE.

4.2 Performance Metrics

The primary performance metric is runtime measured in milliseconds (ms). The comparison metric *Reduction* is computed to represent the relative performance between NemoMap and GK algorithm, and is defined in Equation (3):

$$R_{GK} = \frac{T_{GK} - T_{NemoMap}}{T_{GK}} \times 100 \quad (3)$$

Here, T_{GK} is the runtime of Grochow-Kellis' algorithm in millisecond (*ms*), and $T_{NemoMap}$ is the runtime of NemoMap in *ms*. R_{GK} represents the relative reduction in runtime from using NemoMap instead of GK expressed in percentage.

Another similar metric $R_{ParaMODA}$ is computed to represent the relative performance between NemoMap and ParaMODA, and is defined in the following equation:

$$R_{ParaMODA} = \frac{T_{ParaMODA} - T_{NemoMap}}{T_{ParaMODA}} \times 100 \quad (4)$$

$T_{ParaMODA}$, $T_{NemoMap}$ are the runtime of ParaMODA in *ms* and NemoMap in *ms* respectively, and $R_{ParaMODA}$ is the relative reduction in runtime from using NemoMap instead of ParaMODA expressed in percentage.

4.3 Results and Discussion

We show the results comparing the performance of NemoMap against Grochow-Kellis as well as of NemoMap against ParaMODA in Table 4, Table 5, and Table 6.

4.3.1 NemoMap vs. ParaMODA

In all cases, the implementation of NemoMap ran significantly faster than ParaMODA, and the performance gap widens when the complexity of the network increases. Starting with the simpler network of Hsapi (1,715 nodes and 1,873 edges: edge-to-node ratio of 1.09) in Table 4, we have runtime reduction ranging from 61.29% to 99.57% when using NemoMap; the second network of E. coli has more complexity with a higher edge-to-node ratio (1,223 nodes and 1,654 edges: ratio of 1.35) which results in a higher reduction rate of between 93.58% to 99.97% in runtime as recorded in Table 5; lastly, in Table 6, the most complex network S. cerevisiae with a very high edge-to-node ratio of 1.99 (2,164 nodes and 4,303 edges) yields an even higher reduction range from 94.16% to 99.96% in runtime for NemoMap.

ParaMODA ran out of memory and stopped in several runs, most of which are case with large numbers of mappings (more than 9 million mappings). As discussed in earlier sections, this problem could be addressed by introducing symmetry-breaking technique into the program so that maintaining a list of mappings in memory is not needed. Furthermore, symmetry-breaking also reduced runtime by eliminating redundant isomorphic testing which is the most intensive task of subgraph mapping. Other elements that might help improve runtime include language performance difference between NemoMap and ParaMODA and the level of code optimization.

4.3.2 NemoMap vs. Grochow-Kellis

As mentioned in our previous work [1], performance difference between ParaMODA and GK varied depending on the query graph. Therefore, we attempt to figure out the factors affecting performance difference between NemoMap and GK in this paper by dividing query graphs into groups. For illustration purposes, we sketch the performance in runtime (*ms*) of NemoMap and GK with line charts on the networks of Hsapi, E. coli, and S. cerevisiae in Figure 4, 5, and 6 respectively. Each Figure contains three charts for the three different groups of query graphs as described in Section 4.1: (a), (b), and (c) denotes the Simple, Average, and Complex pattern accordingly.

In most scenarios, GK's algorithm was substantially slower than NemoMap. However, in the case of simple query graphs which had a small count of edges (edge count was smaller than node count), NemoMap runtime was slower as illustrated by chart (a) of Figure 4, 5, and 6 with increasing performance gap when the query graph size gets bigger. Interestingly, NemoMap was faster for simple query graphs of size 4, but it dragged behind for size-5 and larger query graphs. The runtime also got slower when the network becomes more complex: on the Hsapi, NemoMap was behind by 40.79% to 72.06%; for the E. coli, the gap increased to up to 89.30% slowdown; lastly, the S. cerevisiae networks show slower performance of between 49.63% to 90.76%.

On the other hand, when the query graphs became complex with more edges, NemoMap ran much faster than GK. Chart (b) and (c) of Figure 4, 5, and 6 all show exponential improvement in runtime when using NemoMap. Although we did not see an upward trend in performance improvement when the query graph size increases or when the network gets more complex, runtime reduction rate was consistently in the 75% to 90% range as recorded in Table 4, 5, and 6. The only exception was in Table 5 for the Average pattern size-6 query graph with improvement of only 48.84%.

Such observation was consistent in all three tested PPI networks. One possible explanation could be from the isomorphic testing phase, which is the most inten-

Table 4: Performance comparison on the PPI network of Hsapi

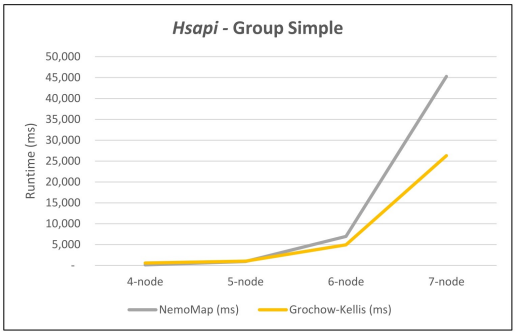
Homo Sapiens (Hsapi): 1,715 nodes - 1,873 edges				
Group Simple	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	3	4	5	6
Number of mappings	32,449	195,061	1,110,425	5,480,057
NemoMap (ms)	122	936	6,951	45,263
Grochow-Kellis (ms)	575	1,002	4,937	26,306
Reduction GK	*78.78%*	*6.59%*	*-40.79%*	*-72.06%*
ParaMODA (ms)	1,531	44,000	1,610,718	N/A
Reduction ParaMODA	*92.03%*	*97.87%*	*99.57%*	*N/A*
Group Average	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	5	6	7	9
Number of mappings	149	1,889	17,686	3,424
NemoMap (ms)	23	50	136	52
Grochow-Kellis (ms)	406	486	914	1,380
Reduction GK	*94.33%*	*89.71%*	*85.12%*	*96.23%*
ParaMODA(ms)	78	484	5,984	12,687
Reduction ParaMODA	*70.51%*	*89.67%*	*97.73%*	*99.59%*
Group Complex	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	6	9	10	13
Number of mappings	17	4	23	1
NemoMap (ms)	24	23	27	35
Grochow-Kellis (ms)	499	475	961	563
Reduction GK	*95.19%*	*95.16%*	*97.19%*	*93.78%*
ParaMODA (ms)	62	78	187	156
Reduction ParaMODA	*61.29%*	*70.51%*	*85.56%*	*77.56%*

Table 5: Performance comparison on the PPI network of E. coli

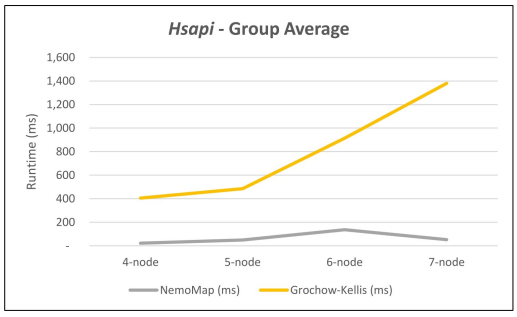
Escherichia coli (E. coli): 1,223 Nodes - 1,654 Edges				
Group Simple	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	3	4	5	6
Number of mappings	91,848	1,072,997	12,494,338	127,746,287
NemoMap (ms)	297	4,531	66,967	879,216
Grochow-Kellis (ms)	830	3,608	31,331	464,460
Reduction GK	*64.22%*	*-25.58%*	*-79.39%*	*-89.30%*
ParaMODA (ms)	5,531	516,593	N/A	N/A
Reduction ParaMODA	*94.63%*	*99.12%*	*N/A*	*N/A*
Group Average	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	5	6	7	9
Number of mappings	606	10,449	95,052	39,750
NemoMap (ms)	12	46	356	199
Grochow-Kellis (ms)	309	1,122	1,395	4,611
Reduction GK	*96.12%*	*95.90%*	*74.48%*	*95.68%*
ParaMODA(ms)	187	3,421	82,125	673,453
Reduction ParaMODA	*93.58%*	*98.66%*	*99.57%*	*99.97%*
Group Complex	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	6	9	10	13
Number of mappings	110	125	1,400	48
NemoMap (ms)	10	14	27	37
Grochow-Kellis (ms)	372	549	1,264	966
Reduction GK	*97.31%*	*97.45%*	*97.86%*	*96.17%*
ParaMODA (ms)	171	953	17,031	112,140
Reduction ParaMODA	*94.15%*	*98.53%*	*99.84%*	*99.97%*

Table 6: Performance comparison on the PPI network of S. cerevisiae

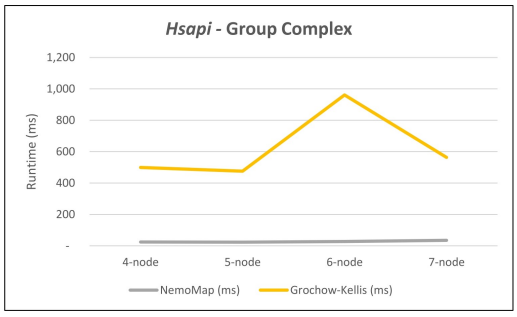
Saccharomyces cerevisiae (S. cerevisiae): 2,164 nodes - 4,303 edges				
Group Simple	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	3	4	5	6
Number of mappings	170,159	1,294,037	9,824,580	67,818,397
NemoMap (ms)	652	6,509	60,693	514,772
Grochow-Kellis (ms)	1,016	4,350	33,507	269,856
Reduction GK	*35.83%*	*-49.63%*	*-81.14%*	*-90.76%*
ParaMODA (ms)	11,171	434,843	N/A	N/A
Reduction ParaMODA	*94.16%*	*98.50%*	*N/A*	*N/A*
Group Average	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	5	6	7	9
Number of mappings	4,869	53,601	327,121	232,404
NemoMap (ms)	79	263	1,453	1,280
Grochow-Kellis (ms)	810	2,291	2,840	9,270
Reduction GK	*90.25%*	*88.52%*	*48.84%*	*86.19%*
ParaMODA(ms)	1,734	37,640	1,289,328	N/A
Reduction ParaMODA	*95.44%*	*99.30%*	*99.89%*	*N/A*
Group Complex	**4-node**	**5-node**	**6-node**	**7-node**
Number of edges	6	9	10	13
Number of mappings	1,591	1,893	22,809	3,904
NemoMap (ms)	59	81	258	437
Grochow-Kellis (ms)	593	803	3,495	2,547
Reduction GK	*90.05%*	*89.91%*	*92.62%*	*82.84%*
ParaMODA (ms)	1,875	20,301	583,343	N/A
Reduction ParaMODA	*96.85%*	*99.60%*	*99.96%*	*N/A*



(a)



(b)



(c)

Figure 4: Graph of runtime against query graph size for (a) Simple graph, (b) Average graph, and (c) Complex graph between NemoMap and Grochow-Kellis for Hsapi. NemoMap outperforms in all cases except for (a).

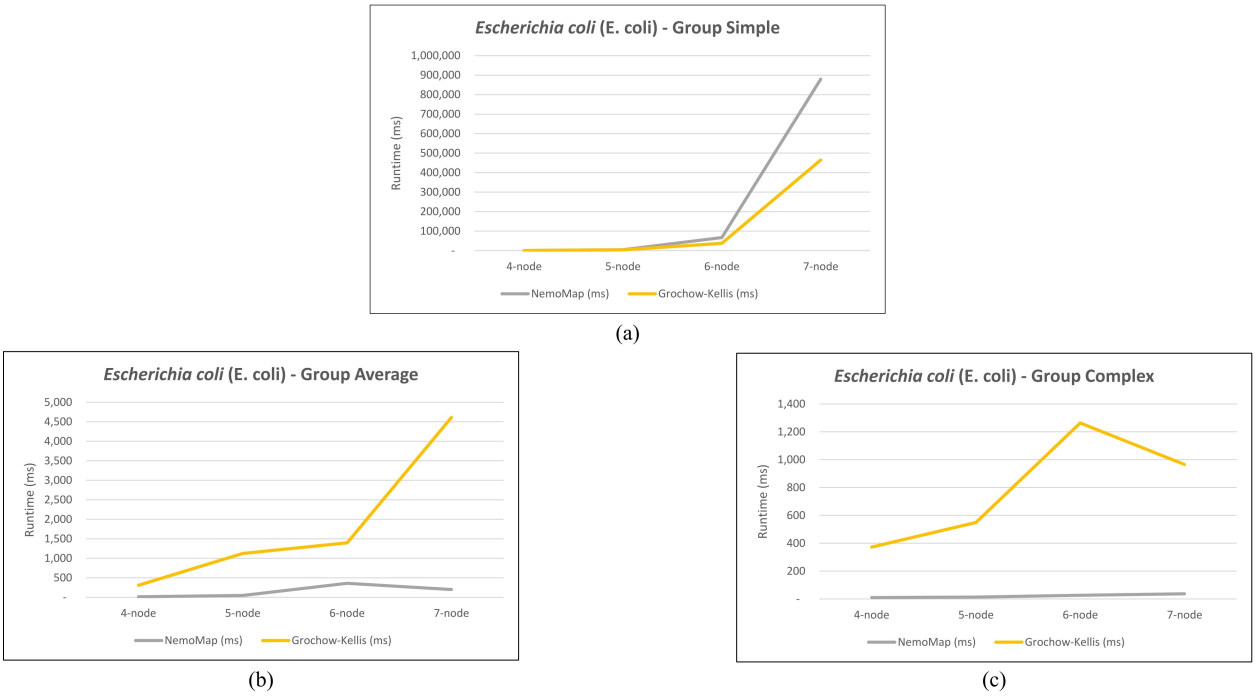


Figure 5: Graph of runtime against query graph size for (a) Simple graph, (b) Average graph, and (c) Complex graph between NemoMap and Grochow-Kellis for E. coli. NemoMap outperforms in all cases except for (a).

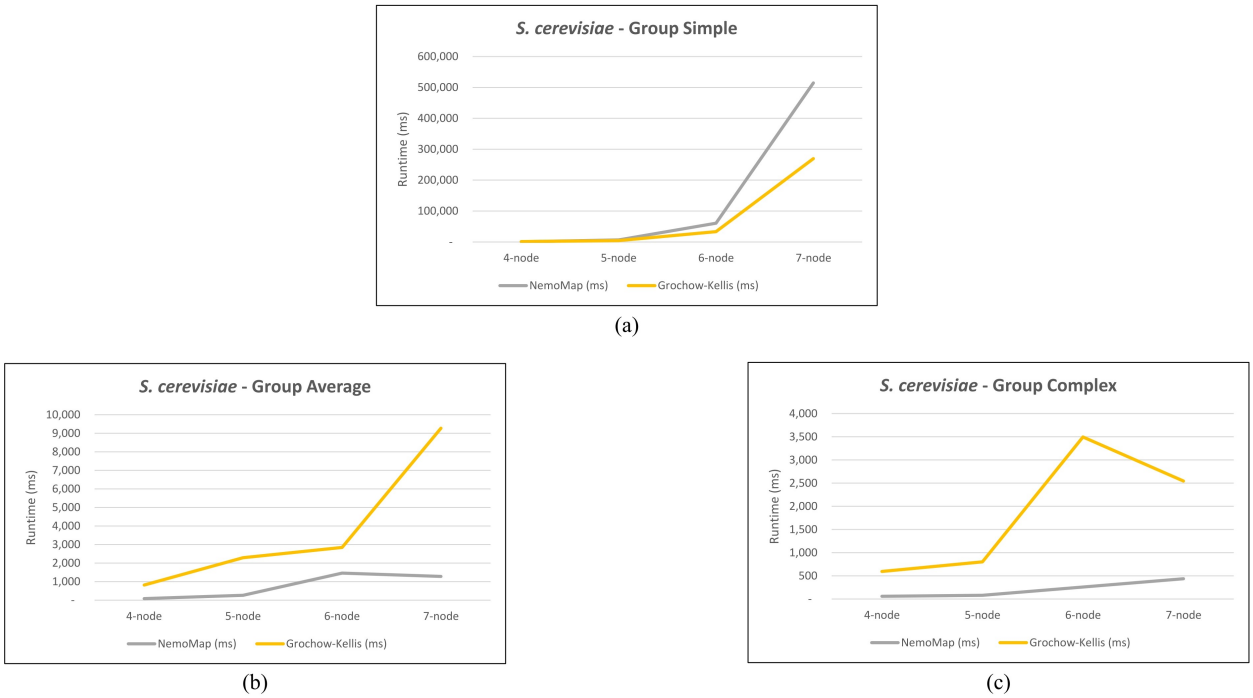


Figure 6: Graph of runtime against query graph size for (a) Simple graph, (b) Average graph, and (c) Complex graph between NemoMap and Grochow-Kellis for S. cerevisiae. NemoMap outperforms in all cases except for (a).

sive task of the subgraph mapping. As discussed in Section 3.1, GK uses two invariants between graphs to test for isomorphism: the degree of each node and the degrees of each nodes neighbors. To clarify the first invariant, when trying to map a node *h* of query graph *H* to a node *g* of network *G*, if node *g*'s degree is smaller than node *h*'s degree then *h* cannot be mapped to *g* and the function must move on to another node *g* as illustrated in Figure 7.

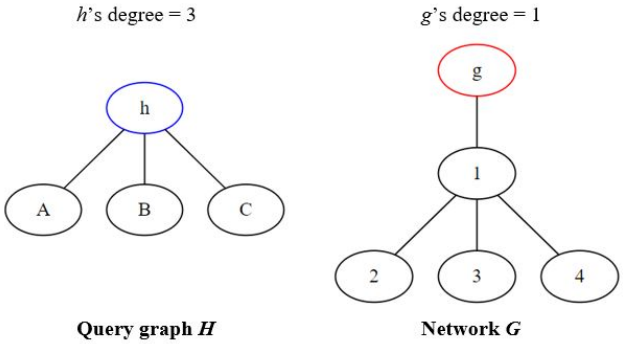
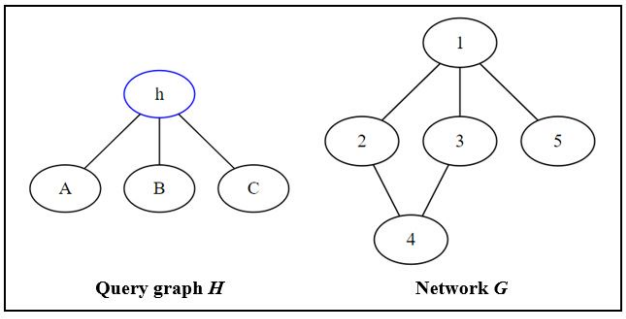


Figure 7: Node *h* cannot be mapped to node g because *g*'s degree is smaller than *h*'s degree



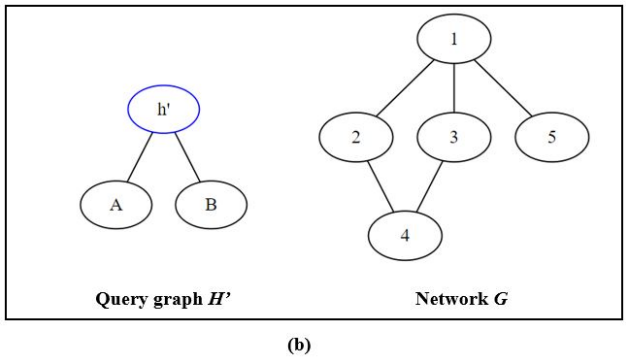


Figure 8: Mapping two different query graphs (a) *H*, and (b) *H'* to the same network *G*

Looking at two different scenarios as illustrated in Figure 8, we have one network *G* and we want to map two query graphs *H* and *H'* to. Assuming that we choose the first mapping node *h* and *h'* as highlighted for each query graph, and we examine the possibility of mapping node *h* to each node of network *G*.

In scenario (a), node *h* has a degree of three so only node 1 of *G* can support it. As a result, the recursive call for isomorphic testing will only be called and branched out when trying to map node h to 1; node 2, 3, 4, and 5 will be aborted early so no isomorphic test is ran. On the other hand, in scenario (b), node *h'* has a degree of two so node 1, 2, 3, and 4 can all support it, resulting in a lot more of isomorphism recursive tests since only mapping to node 5 can be aborted early.

The larger number of isomorphic tests and the lower chance to abort early in (b) are advantageous for GK because it uses the same isomorphic testing algorithm as NemoMap, but the network is getting smaller and smaller after every iteration because GK removes mapped nodes from the network unlike the fixed network of NemoMap. However, GK needs to iterate over all nodes of *H'*, so the removal effect might get canceled out. If the query graph is highly symmetrical and symmetry-breaking is used, then it will unlikely to be fully even out. Recall that with symmetry-breaking, GK only iterates over representative nodes of equivalence class. In (b), node *A*, *B*, and *C* belong to the same equivalence class as they are symmetrical (for more detail on equivalence class please reference [10]), so instead of iterating over four nodes (*h'*, *A*, *B*, and *C*), Grochow-Kellis would only iterate over two nodes (*h'* and one arbitrary node from the equivalence class). Such effect might help GK retain better performance on simpler query graphs with lower nodes' degrees and large equivalence classes.

As mentioned in the previous section, other elements that might affect runtime include language performance difference between C++ and Java and the level of code optimization.

5 Conlusion and Future Works

In this paper, we present NemoMap, a motif-centric algorithm inspired by the work of Grochow and Kellis [10], as a tool to determine the frequency of a subgraph's appearance in a target network. Extending from our previous work in ParaMODA [1], the new experimental results show that there is improvement in performance using NemoMap, yet it does not apply in all cases. Specifically, we observe that NemoMap runs much faster when analyzing complex query graphs with more edges, while Grochow-Kellis' solution is more efficient for simpler query graphs with fewer edges (i.e. lower degrees) and high symmetry. As an extension, NemoMap also sees large improvement in runtime and memory usage compared to the predecessor ParaMODA due to the utilization of symmetry-breaking technique and the better node selection process.

Additionally, NemoMap structure allows for an easier parallelization scheme, as multiple query graphs or different nodes of the same query graph can be assigned to different processors running concurrently. Such operation is necessary for future research as the need for finding larger network motif becomes more prominent. Another essential future item is the implementation of a random network generator and a

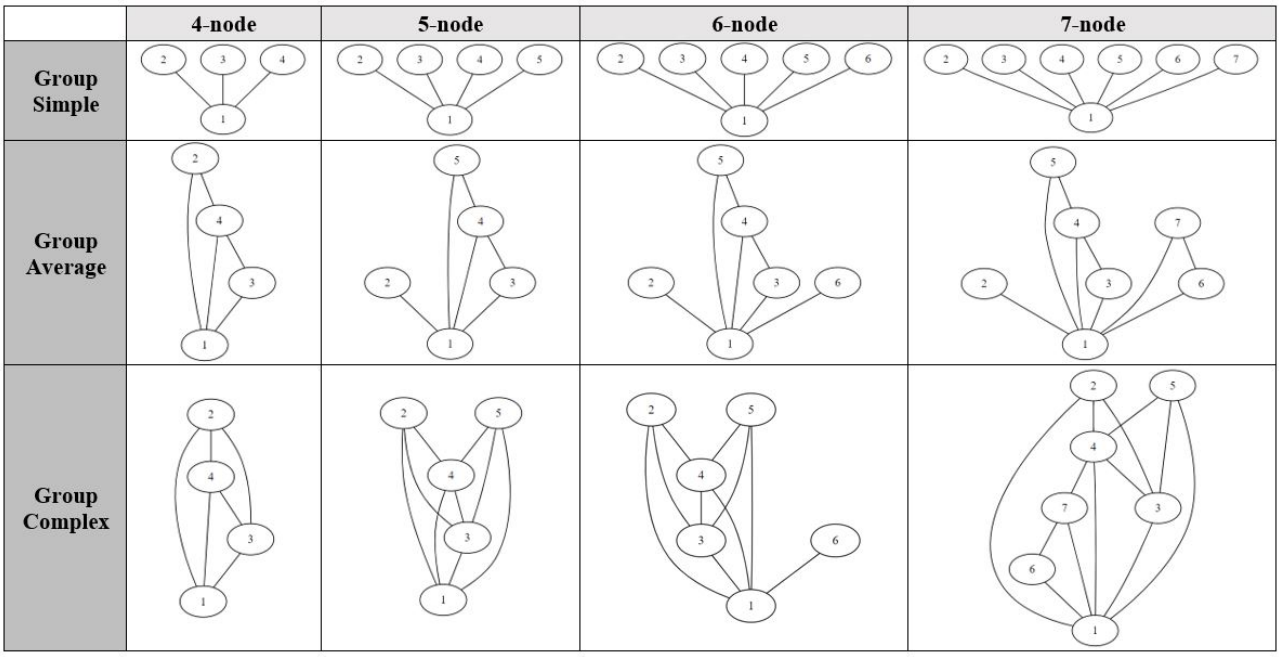


Figure 9: Visualization of groups of query graphs used for testing

statistical analysis system into the mapping function so that the NemoMap program could become a complete network motif detection suite.

The source code of NemoMap and usage instruction are available on https://github.com/tienhuynh/NemoMap. ParaMODA is also available on https://github.com/smbadiwe/ParaMODA.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We thank the Computing and Software Systems, School of STEM, at the University of Washington Bothell for the support of this research.

References

[1] S. Mbadiwe and W. Kim. Paramoda: Improving motif-centric subgraph pattern search in ppi networks. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1723–1730, Nov 2017.

[2] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, March 05 2004.

[3] Manuel Middendorf, Etay Ziv, and Chris H. Wiggins. Inferring network mechanisms: The drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3192–3197, 2005.

[4] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[5] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon. On the uniform generation of random graphs with prescribed degree sequences. *eprint arXiv:cond-mat/0312028*, December 2003.

[6] I. Albert and R. Albert. Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352, December 2004.

[7] Yuji Zhang, Jianhua Xuan, Benilo G de los Reyes, Robert Clarke, and Habtom W. Ressom. Network motif-based identification of breast cancer susceptibility genes. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5696–5699. IEEE, aug 2008.

[8] Bjvrn H. Junker and Falk Schreiber. *Analysis of Biological Networks*. Wiley, 2008.

[9] W. Kim, M. Diko, and K. Rawson. Network motif detection: Algorithms, parallel and cloud computing, and related tools. *Tsinghua Science and Technology*, 18(5):469–489, Oct 2013.

[10] Joshua A. Grochow and Manolis Kellis. Network motif discovery using subgraph enumeration and symmetry-breaking. In *Proceedings of the 11th annual international conference on Research in computational molecular biology*, RECOMB'07, pages 92–106, Berlin, Heidelberg, 2007. Springer-Verlag.

[11] Andrew Andersen and Wooyoung Kim. *NemoLib: A Java Library for Efficient Network Motif Detection*, pages 403–407. Bioinformatics Research and Applications: 13th International Symposium, ISBRA 2017, Honolulu, HI, USA, May 29 June 2, 2017, Proceedings. Springer International Publishing, Cham, 2017.

[12] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in Bioinformatics*, 2011.

[13] B. McKay. Practical graph isomorphism. *Congr Numer*, 30:45–87, 1981.

[14] F. Schreiber and H. Schwobbermeyer. Mavisto: a tool for the exploration of network motifs. *Bioinformatics*, 21:3572–3574, 2005.

[15] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating sub-graph concentrations and detecting network motifs. *Bioinformatics*, 20:1746–1758, 2004.

[16] S. Wernicke. Efficient detection of network motifs. *IEEE/ACM transactions on computational biology and bioinformatics*, 3(4):347–359, December 01 2006.

[17] S. Wernicke and F. Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22:1152–1153, 2006.

[18] Zahra Kashani, Hayedeh Ahrabian, Elahe Elahi, Abbas Nowzari-Dalini, Elnaz Ansari, Sahar Asadi, Shahin Mohammadi, Falk Schreiber, and Ali Masoudi-Nejad. Kavosh: a new algorithm for finding network motifs. *BMC Bioinformatics*, 10(1):318, 2009.

[19] Jin Chen, Wynne Hsu, Mong Li Lee, and See-Kiong Ng. Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 106–115, New York, NY, USA, 2006. ACM.

[20] Xin Li, Douglas S. Stones, Haidong Wang, Hualiang Deng, Xiaoguang Liu, and Gang Wang. NetMODE: Network motif detection without nauty. *PLoS ONE*, 7(12):e50093, 2012.

[21] Saeed Omidi, Falk Schreiber, and Ali Masoudi-Nejad. Moda: An efficient algorithm for network motif discovery in biological networks. *Genes and Genetic Systems*, 84(5):385–395, 2009.

[22] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S. M. Kim, and D. Eisenberg. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305, Jan 1 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 193-200 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Design of smart chess board that can predict the next position based on FPGA

Alaa Hamza Omran*,1, Yaser Muhammad Abid2

[1]*Department of Information System Management, University of Information Technology and Communication, Baghdad, 00964, Iraq*

[2]*Department of Business Information Technology, University of Information Technology and Communication, Baghdad, 00964, Iraq*

A R T I C L E I N F O

Article history:
Received: 29 June, 2018
Accepted: 26 July, 2018
Online: 29 July, 2018

Keywords:
Artificial Neural Network
Smart Chessboard
Back propagation algorithms
Chess game, intelligent chessboard
Intelligent controller and Supervised feed forward algorithm

A B S T R A C T

The abilities of human brain to discover solutions for many problems is a great gift that motivate the scientists to develop the revolution of the artificial intelligence and using it in many areas. This paper proposed an intelligent chessboard which works in a way that similar to the human brain that predicts the next positions for any piece of the chess. MATLAB is used in the training of the artificial neural networks of the chessboard and implemented on FPGA as a hardware part; the proposed system has many advantages such as the low cost of production with a high efficiency in the prediction of the position of any piece on the chessboard and the similarities to the human brain.

1. Introduction

This paper is an extension of work originally presented in 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS) [1].

In the early centuries, wars among major countries were in the zenith of prosperity, and the countries that have consultants who develop strategic plans for the Wars were prevail these wars; therefore, a game known as "kings game" was appeared. This game helps in developing the strategic plans without any human losses; it was spread rapidly in most countries and enticing many people because of its ability to motivate the human mind to think in a smart way. Furthermore, many schools in the world start to teach this game and considering it as the one of the games that test the intelligence [2]. Currently, and due to the tremendous evolution of the modern technology, many researches tries to make this game plays on the computers [3] [4]. They start to develop many programs that simulate chess game, however, playing this game can take many hours as known and sitting on computers for many hours can leads to serious problem such as headache, blurred vision and back pain etc. a large number of papers proved that the neural networks and intelligent systems have great dependency in many areas [5] [6] [7]. Many attempts in papers to build intelligent chess games have been done with disadvantages of very high costs used for designing the systems and complex electrical circuit.

However, there are a great number of researches in this area; [8] presents chess game in real time by using MATLAB environment without using for any chess engine, the system designed to detecting the chess movements depending on RGB webcam, and that means the complexity in the system designed may obtain incorrect positions during the game, the system have high cost with low accuracy. [9] presents chess board that can play against human depending on robotic arm which is five degree of freedom, the control system designed to control the speed and the position of the robotic arm, the designed system depend on the use of computer and that too expensive with very complex in the design that take long period of time to control the speed and the position. [10] Presents an attempt to design chess play against human and calculate its moves, the system implemented on microcontroller and the computer used to control chess movement, the system gave not perfect results and the use of computer also

* Corresponding author: Alaa Hamza Omran, Email: engineeralaa90@gmail.com

have disadvantages as mentioned. [11] Presents a robotic system play chines chess with human that can recognize pieces on the designed chessboard, the system designed to move the pieces of chess with the help of mechanical arm of robot, the system designed with expensive components. [12] Presents the chess game designed using MATLAB and electronics circuit for educational reasons, the system implemented on Arduino and the detection process for any piece to his place and type done by using Infrared light-emitting diode and photodiode, extra multiplexers, encoders and shift registers to increase the no. of I/P or O/P pins, this system have complex designed electronics circuit and high cost. [13] [14] present the FPGA in different systems and producing great efficiency in different area of researches and high speed in response. This paper merges two methods; the first one, is the ordinary method which used in the playing of the chess game. And the second method is the newest that used the technology methods to facilitate the playing of the game without using any computer device to decrease the damage that can be caused from playing the game as much as possible. The main contribution of this paper is to design an intelligent chessboard that can present the next positions for each piece and the avoidance of using computer due to different disadvantages that can occurred as a results of using it such as headache and blurred vision etc. The chessboard is designed by using a simple and low cost of electronics components that can compensate the using of computers which are used in many other chess game application.

2. Proposed System

Intelligent chessboard that can predict the next positions for all pieces of chess game is designed in this paper. Figure (1) shows the block diagram of the proposed system.

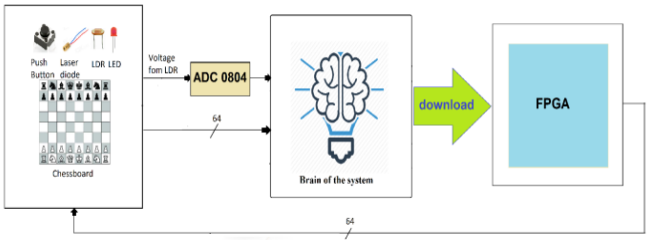


Figure 1: Block diagram of intelligent chessboard system

It consists of 64 push buttons which are equally distributed, one button for each square on the chessboard; a simple press on any push button of the intelligent chessboard would send a notification to the proposed system with the exact position and the needing to predict the next positions. There are also 64 LASER diodes which are equally distributed as the push buttons, one LED for each square of the chessboard. In the chess game, there are 32 pieces which are divided equally into 16 pieces, (king, queen, rooks, Bishop, knights and pawns), for each player. Therefore, 32 LDR sensors were used and equally distributed as one sensor for each piece. Furthermore, the pressing of any piece would make the LEDs of chessboard glowed in order to notify the user with the next positions were the pressed piece can move on. The LDR sensors were used to distinct among the 32 pieces; they are connected serially with a constant resistance for each piece as that shown in Figure (2)

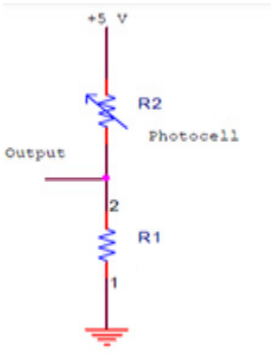


Figure 2: Constant resistance connected with LDR sensor

The light falling on the LDR sensor will produces a fixed range of voltage for each piece; because for each piece there is a different constant resistance; table (1) shows chess's piece types with corresponding voltages in the proposed system.

Table 1: Chess's piece types with corresponding voltages

Piece type	Range of voltage
Rook pieces	(from 0 to 0.5) volt
Knight pieces	(From 0.6 to 1)volt
Bishop pieces	(From 1.1 to 1.5) volt
King pieces	(From 1.6 to 2) volt
Queen pieces	(from 2.1to 2.5) volt
White pawn pieces	From (2.6 to 3) volt
black pawn pieces	(from 3.1to 3.5) volt

In the chess game, there are 8 pawns for each player which can move in different paths and opposite direction; therefore, the range of voltages differ and divided into two team white pawns and black pawns as shown in table (1).

3. Simulation Results of Training

The human brain handles the processes in an amazing and smart way that distinct it from any other machines. Therefore, the neural networks of the proposed system trained in such a way that is similar to the human brain in order to notify the players with the next positions of the movement for every piece in a smart way. Super vised Feed forward neural networks were used to train all the pieces of chess game depending on back propagation algorithm [15]; and each piece types in chess was trained individually. The determination of the No. of neurons in the input layers, No .of neurons in the hidden layers, the No. of layers and also the activation functions in each network all done using Try and error method.

3.1. Rook

Finally, the rook piece in the chess game moves in different paths as compared with the other pieces, it has specific paths of movements on chessboard which can moves horizontally or vertically. Four no. of rook pieces in chess game which are divided equally into two teams, all rooks piece have the same paths of movements on the chessboard; therefore, the training of the rook

was done individually for the first piece and repeated to the other pieces. In order to train the neural networks with the path of movements of the rook piece to act such as human brain, every square of the chessboard has sequential number begins with 1 and ends with 64. Any square which carry the rook piece will predict a different probability of the movements for the next positions of rook. Therefore, the networks of the chessboard trained with all possibility of movements for the rook piece on the chessboard. They are trained with six neurons in the hidden layer and fifteen neurons in the output layer as shown in Figure (3)

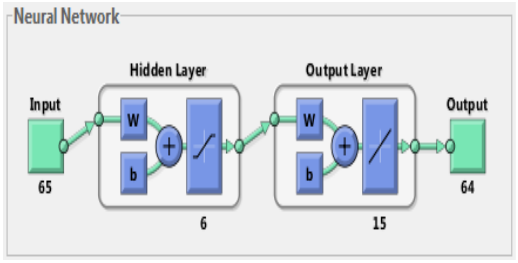


Figure 3: Final neural network used for rook pieces

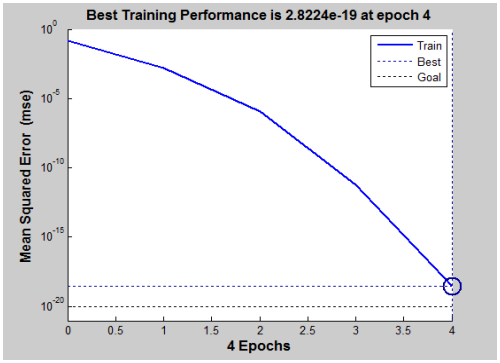


Figure 4: The simulation results of training the rook pieces

Figure (4) shows the simulation results of training the rook pieces; if the square with the number one carry the rook piece, all the several paths for the next position of the rook piece are shown in figure (5).

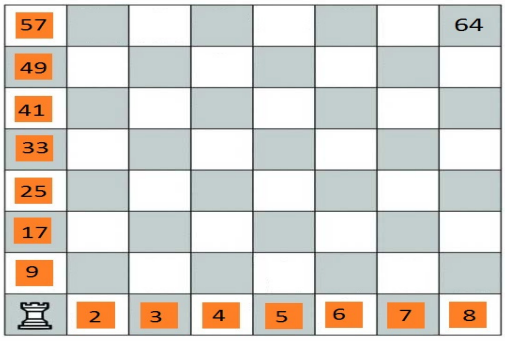


Figure 5: The way of movements of rook piece on the chessboard.

In the proposed chessboard a simple press over the rook piece, will lead to press the push button at the square with number one at the chessboard which will activate the laser diode to fall its light on the LDR sensor and depending on the voltage which produced from the LDR sensor and according to the range of voltage specified previously in table (1), the rook piece will be recognized. In more details, the reorganization process will lead to change the state of the LEDs to ON state over the designed chessboard and the exact squares (2, 3, 4, 5, 6, 7, 8, 9, 17, 25, 33, 41, 49, and 57) will be glowed to notify the player with the next available positions that the rook can move on.

The trained networks of rook piece converted into Simulink in order to test it and connect it with other networks in the next steps. The simulation results of rook piece movement for next positions shown in figure (6)

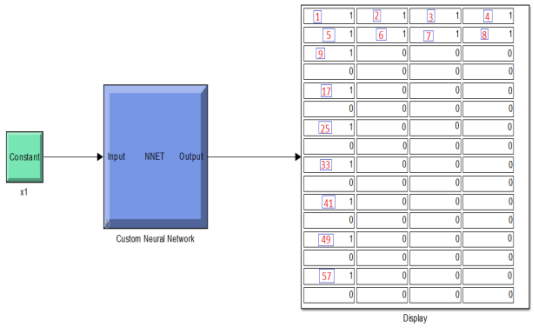


Figure 6: Simulation results of rook pieces for next positions movements

The trained networks of the rook pieces converted into VHDL code by using MATLAB that have HDL coder in order to implement it on FPGA, figure (7) shows the RTL of rook pieces used in proposed system and figure (8) shows the simulation results of intelligent networks of rook by using Xilinx ISE Design Suite 13.3.

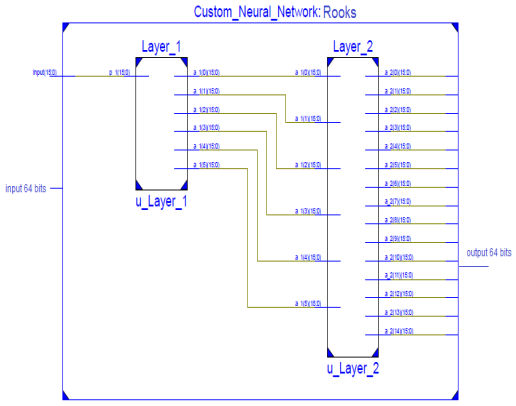


Figure 7: The RTL of rook pieces used in proposed system

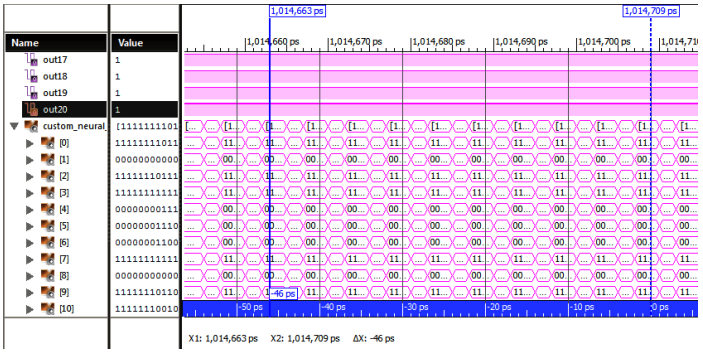


Figure 8: The simulation results of intelligent networks of rook

3.2. Bishop

The Bishop piece in the chess game moves in different paths as compared with the other pieces, it has specific paths of

movements on chessboard which can moves diagonally. The chess game has four pieces of bishop which are divided equally into two teams, all bishops piece have the same paths of movements on the chessboard; therefore, the training of the bishop was done individually for the first piece and repeated to the other pieces. In order to train the neural networks with the path of movements of the bishop piece to act such as human brain, every square of the chessboard has sequential number begins with 1 and ends with 64. Any square which carry the bishop piece will predict a different probability of the movements for the next positions of bishop. Therefore, the networks of the chessboard trained with the all possible movements of the bishop piece on the chessboard. They are trained with eight neurons in the hidden layer and twenty neurons in the output layer as shown in Figure (9).

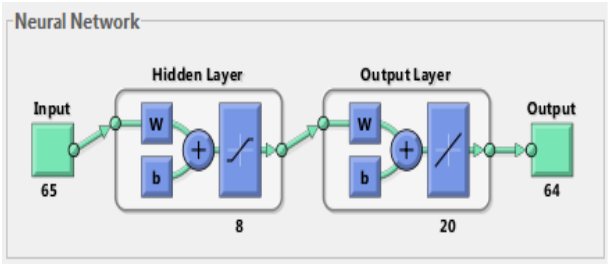


Figure 9: Final neural network used for bishop pieces

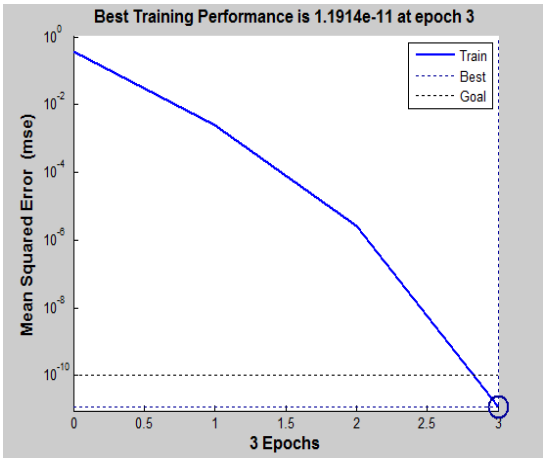


Figure 10: The simulation results of training bishop pieces

Figure (10) shows the simulation results of training bishops; if the square with the number twenty two carry the bishop piece, all the several paths for the next position of the bishop piece are shown in figure (11).

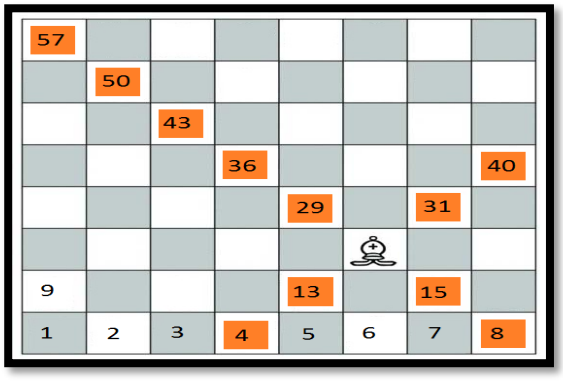


Figure 11: The way of movements of bishop pieces on the chessboard.

In the proposed chessboard a simple press over the bishop piece at square twenty two, will lead the same processes mentioned in the rook piece and the bishop piece will be recognized. In more details, the reorganization process will lead to change the state of the LEDs to ON state over the designed chessboard and the exact squares (4, 8, 13, 15, 22, 29, 31, 36, 40, 43, 50, and 57) will be glowed to notify the player with the next available positions that the bishop can move on.

The trained networks of bishop piece converted into Simulink in order to test it and connect it with other networks in the next steps. Figure (12) show the simulation results of bishop piece movement for next positions.

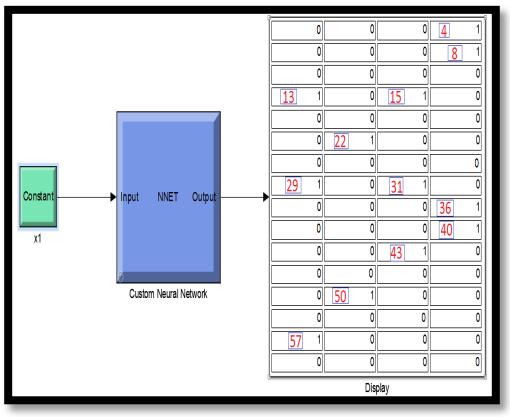


Figure 12: Simulation results of bishop pieces for next positions movements

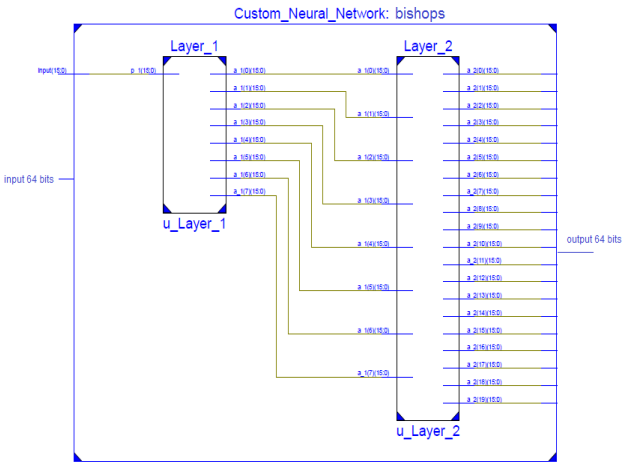


Figure 13: The RTL of bishop pieces used in proposed system

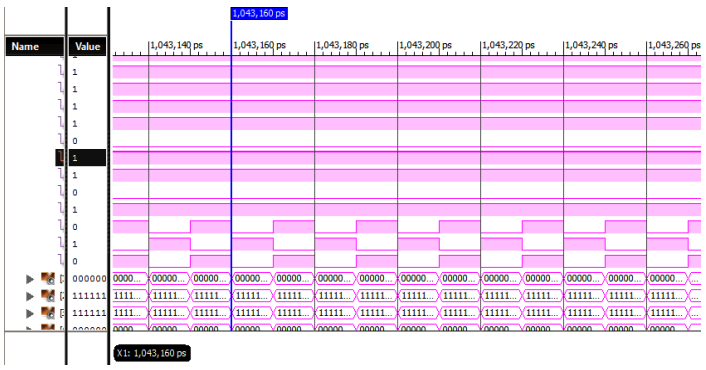


Figure 14: The simulation results of intelligent networks of bishop

The trained networks of the bishop pieces converted into VHDL code by using MATLAB that have HDL coder in order to implement it on FPGA, figure (13) shows the RTL of bishop pieces used in proposed system and figure (14) shows the simulation results of intelligent networks of bishop by using Xilinx ISE Design Suite 13.3.

3.3. Knights

The knight piece in the chess game moves in different paths as compared with the other pieces, it has specific paths of movements on chessboard which can moves in the shape of **L** for two square and can jumps on other pieces on the chessboard. The chess game has four pieces of knights which are divided equally into two teams, all knights piece have the same paths of movements on the chessboard; therefore, the training of the knight was done individually for the first piece and repeated to the other pieces. In order to train the neural networks with the path of movements of the knight piece to act such as human brain, every square of the chessboard has sequential number begins with 1 and ends with 64. Any square which carry the knight piece will predict a different probability of the movements for the next positions of knight. Therefore, the networks of the chessboard trained with the all possible movements of the knight piece on the chessboard. They are trained with twelve neurons in the hidden layer and eleven neurons in the output layer as shown in Fig (15).

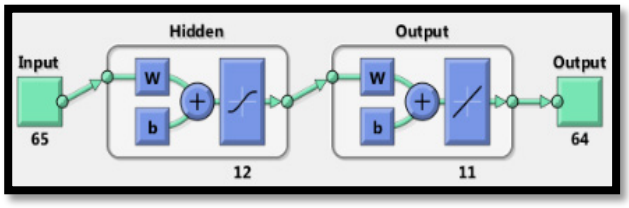


Figure 15: Final neural network used for knight pieces

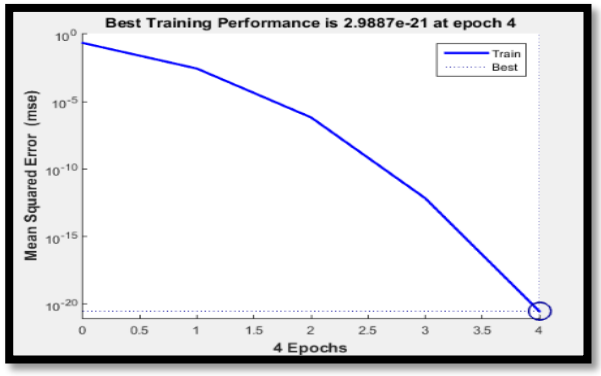


Figure 16: The simulation results of training the knight pieces

Figure (16) shows the simulation results of the knight movements; if the square with the number thirty seven carry the knight piece, all the several paths for the next position of the knight piece are shown in figure (17).

In the proposed chessboard a simple press over the knight piece at square thirty seven, will lead the same processes mentioned in the rook piece and the knight piece will be recognized. In more details, the reorganization process will lead to change the state of the LEDs to ON state over the designed chessboard and the exact squares (20, 22, 27, 31, 37, 43, 47, 52, and 54) will be glowed to notify the player with the next available positions that the knight can move on.

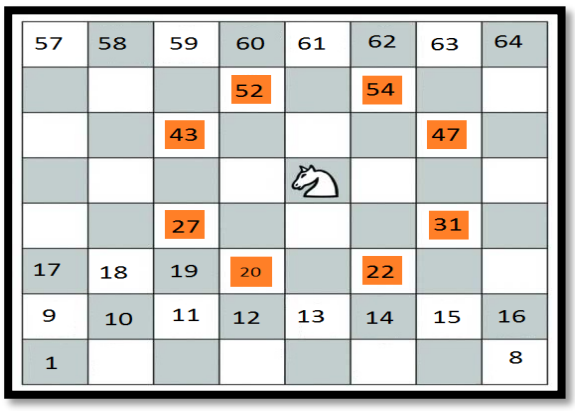


Figure 17: The way of movements knight pieces on the chessboard.

The trained networks of knight piece converted into Simulink in order to test it and connect it with other networks in the next steps. The simulation results of knight piece movement for next positions shown in figure (18)

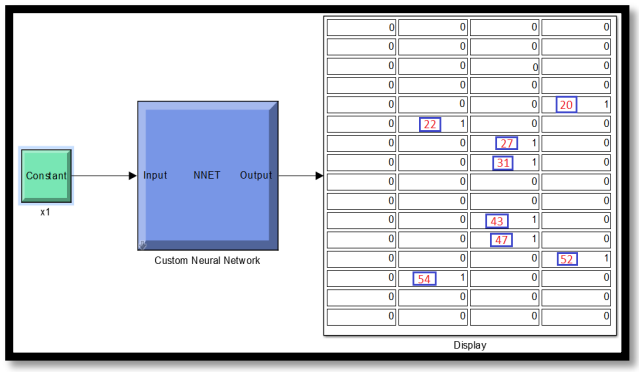


Figure 18: Simulation results of knight pieces for next positions movements

The trained networks of the knight pieces converted into VHDL code by using MATLAB that have HDL coder in order to implement it on FPGA, figure (19) shows the RTL of knight pieces used in proposed system and figure (20) shows the simulation results of intelligent networks of knight by using Xilinx ISE Design Suite 13.3.

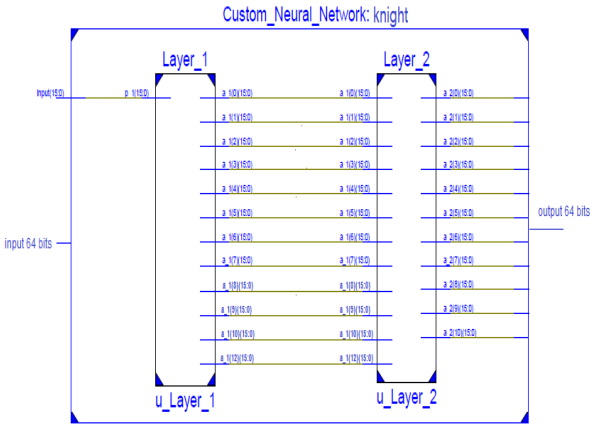


Figure 19: The RTL of knight pieces used in proposed system

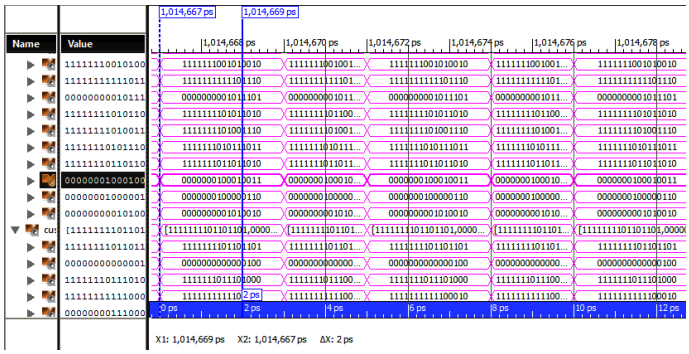


Figure 20: The simulation results of intelligent networks of knight

3.4. Queens

The queen piece in the chess game moves in different paths as compared with the other pieces, it is the strongest piece and has specific paths of movements on the chessboard, it can moves at paths that collect the both paths of the rook and bishop pieces at the chessboard. The chess game has two pieces of queen which are divided equally into two teams, all queen pieces have the same paths of movements on the chessboard; therefore, the training of the queen was done individually for the first piece and repeated to the other piece. In order to train the neural networks with the path of movements of the queen piece to act such as human brain, every square of the chessboard has sequential number begins with 1 and ends with 64. Any square which carry the queen piece will predict a different probability of the movements for the next positions of queen. Therefore, the networks of the chessboard trained with the all possible movements of the queen piece on the chessboard. They are trained with four neurons in the hidden layer and twenty neurons in the output layer as shown in Figure (21).

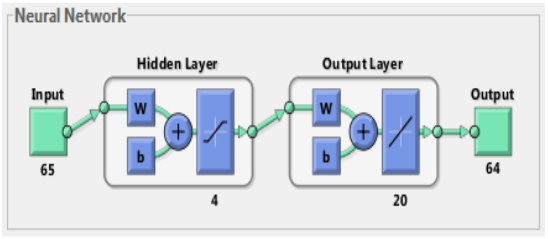


Figure 21: Final neural network used for queen pieces

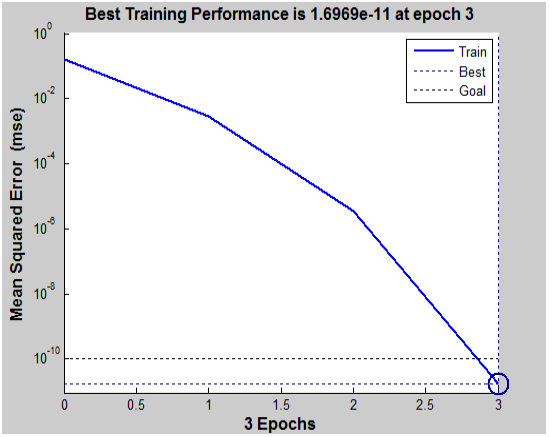


Figure 22: The simulation results of training queen pieces

Figure (22) shows the simulation results of training the queen pieces; if the square with the number thirty three carry the queen piece, all the several paths for the next position of the queen piece are shown in figure (23).

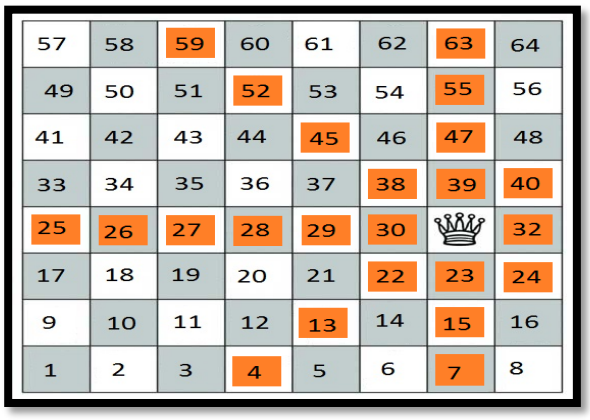


Figure 23: The way of movements queen pieces on the chessboard.

In the proposed chessboard a simple press over the queen piece at square thirty one, will lead the same processes mentioned in the rook piece and the queen piece will be recognized. In more details, the reorganization process will lead to change the state of the LEDs to ON state over the designed chessboard and the exact squares (4, 7, 13, 15, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 38, 39, 40, 45, 47, 52, 55, 59, 63) will be glowed to notify the player with the next available positions that the queen can move on.

The trained networks of knight piece converted into Simulink in order to test it and connect it with other networks in the next steps. The simulation results of queen piece movement for next positions shown in figure (24)

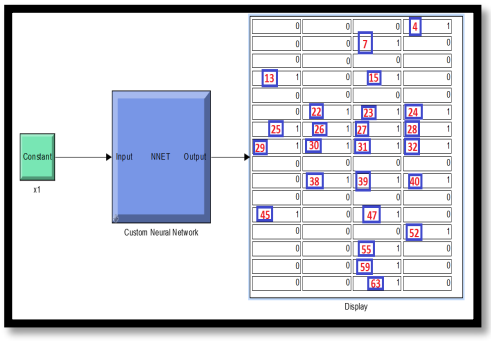


Figure 24: Simulation results of queen pieces for next positions movements

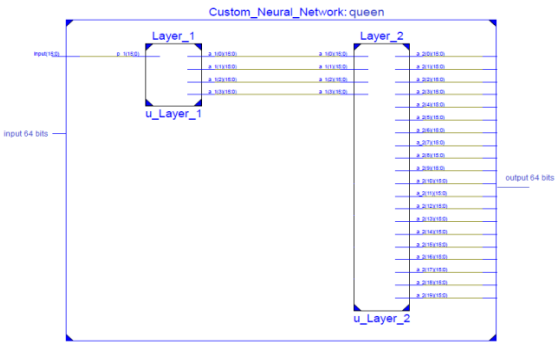


Figure 25: The RTL of queen pieces used in proposed system

The trained networks of the queen pieces converted into VHDL code by using MATLAB that have HDL coder in order to implement it on FPGA, figure (25) shows the RTL of queen pieces used in proposed system and figure (26) shows the simulation results of intelligent networks of queen by using Xilinx ISE Design Suite 13.3.

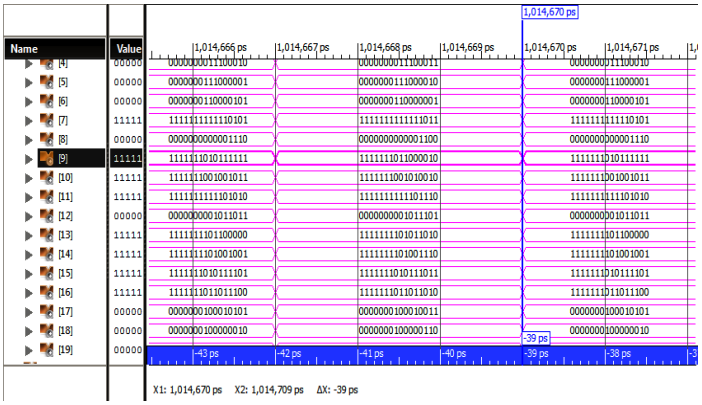


Figure 26: The simulation results of intelligent networks of queen

3.5. Kings

The king piece in the chess game moves in different paths as compared with the other pieces, it has specific paths of movements on the chessboard, it moves such as the queen piece with only one square. The chess game has two pieces of king which are divided equally into two teams, all king pieces have the same paths of movements on the chessboard; therefore, the training of the king was done individually for the first piece and repeated to the other piece.

In order to train the neural networks with the path of movements of the king piece to act such as human brain, every square of the chessboard has sequential number begins with 1 and ends with 64. Any square which carry the king piece will predict a different probability of the movements for the next positions of king. Therefore, the networks of the chessboard trained with the all possible movements of the king piece on the chessboard. They are trained with nine neurons in the hidden layer and twenty neurons in the output layer as shown in Figure (27).

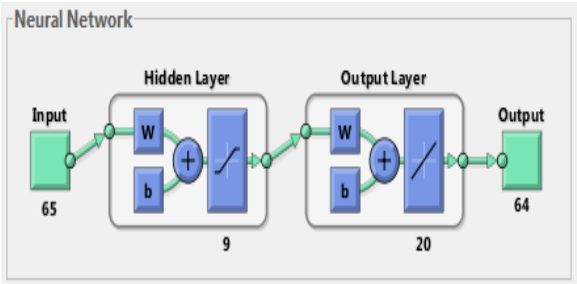


Figure 27: Final neural network used for king pieces

Figure (28) shows the simulation results of training the king pieces; if the square with the number forty-three carry the king piece, all the several paths for the next position of the king piece are shown in figure (29).

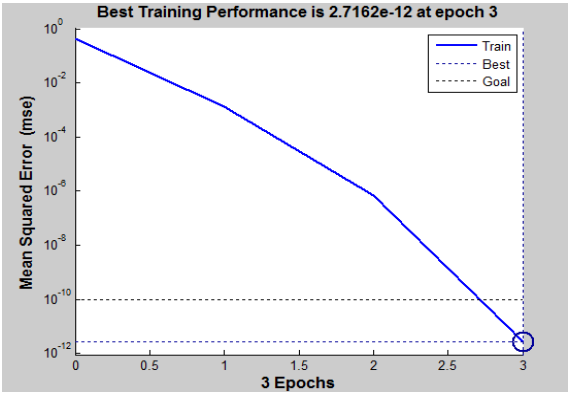


Figure 28: The simulation results of training the king pieces

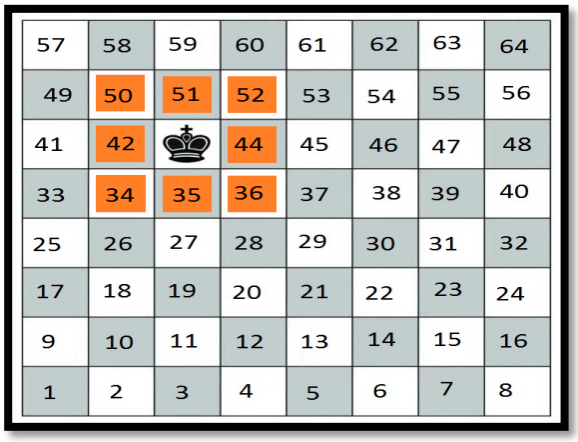


Figure 29: The way of movements of the king pieces on the chessboard.

In the proposed chessboard a simple press over the king piece at square forty three, will lead the same processes mentioned in the rook piece and the king piece will be recognized. In more details, the reorganization process will lead to change the state of the LEDs to ON state over the designed chessboard and the exact squares (34, 35, 36, 42, 43, 44, 50, 51 and 52) will be glowed to notify the player with the next available positions that the king can move on.

The trained networks of king piece converted into Simulink in order to test it and connect it with other networks in the next steps. The simulation results of king piece movement for next positions shown in figure (30)

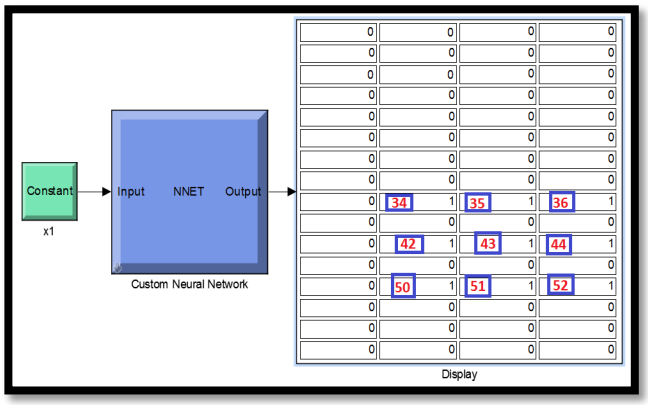


Figure 30: Simulation results of king pieces for next positions movements

The trained networks of the king pieces converted into VHDL code by using MATLAB that have HDL coder in order to implement it on FPGA, figure (31) show the RTL of king pieces used in proposed system and figure (32) shows the simulation results of intelligent networks of king by using Xilinx ISE Design Suite 13.3.

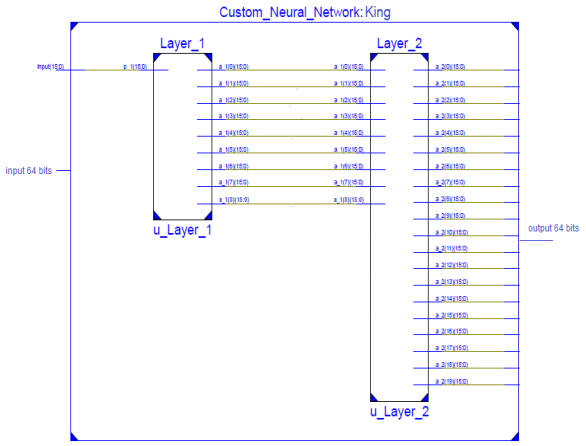


Figure 31: The RTL of king pieces used in proposed system

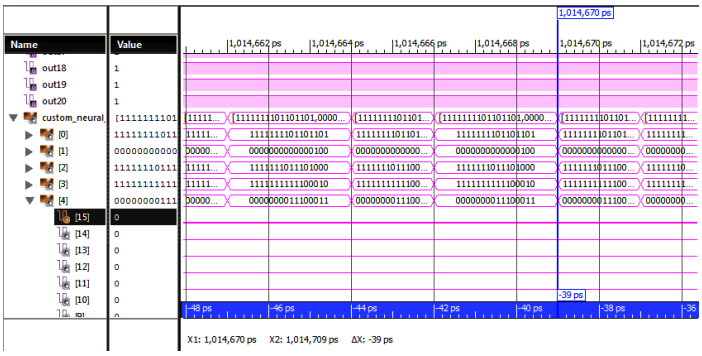


Figure 32: The simulation results of intelligent networks of king

3.6. Pawns

The movement of pawn piece was the simplest and the easiest to know as compared with the other pieces as well as to reduce the cost of the internal design, whereas knowing the movement of the pawn piece was just single step either to the straight ahead or diagonally one step in the case of attack. Therefore, it was not mentioned with the other scenarios.

4. Conclusion

In this paper an intelligent chessboard designed to handle the processes such as human brain through predict the next positions of any piece at the chessboard. Different algorithms used in the training of the chessboard and the best one were used in the proposed system. Each piece of chess game trained individually by using MATLAB and the results tested with different states, all networks of pieces trained was combined together in one system that produce the intelligent chessboard. It converted into VHDL code by using HDL coder in MATLAB, the VHDL code of the proposed system tested with different states by using ISE Design Suite 13.3.

References

[1] Alaa Hamza Omran, Yaser M. Abid, Dr. Huda Kadhim, "Design of Artificial Neural Networks System for Intelligent Chessboard", 4th IEEE international conference on engineering technology and applied scince (ICETAS), Manama, Bahrain, NOV. 2017

[2] Manish A. Chhangani, " Arduino based Wireless Powered Chess ", International Journal of Innovative Research in Computer and Communication Engineering,, Vol. 3, Issue 4, April 2015.

[3] Wang LinLin, Wu YuNu, Wang YaJie, Cao Guoqiang, "Research on the Surakarta chess game program based on Unity 3D", IEEE, 2017.

[4] NasrulHumaimi Mahmood1, Che Ku MohdSalahuddinChe Ku Long, IsmawatiAbdGhani, RubitaSudirman, "Low Cost Electronic Chess Set for Chess Tournament", IEEE 7th International Colloquium on Signal Processing and its Applications, 2011.

[5] Diwas Sharma, Udit Kr. Chakraborty, "An Improved Chess Machine based on Artificial Neural Networks", International Journal of Computer Applications® (IJCA) (0975 – 8887) National Conference cum Workshop on Bioinformatics and Computational Biology, NCWBCB- 2014.

[6] Panos J. Antsaklis, "Neural Networks in Control Systems", IEEE Control Systems M agazine , Vol.10, No.3, pp.3-5, April 1990; Special Issue on 'Neural Networks in Control Systems' of the I EEE Control Systems Magazine , Vol.10, No.3, pp.3-87, April 1990.

[7] Kuo-Lan Su, Bo-Yi Li, Jr-Hung Guo, Kuo-Hsien Hsia, "Implememtation of The Chess Game Artificial Intelligent Using Mobile Robots",SCIS&ISIS 2014, Kitakyushu, Japan, December 3-6, 2014.

[8] Can Koray , Emre Sumer, "A Computer Vision System for Chess Game Tracking", 21st Computer Vision Winter Workshop Luka Cehovin, RokMandeljc, Vitomir ˇ Struc (eds.) ˇ RimskeToplice, Slovenia, February 3–5, 2016

[9] Firas Abdullah Thweny Al-Saedi , Ali H. Mohammed, "Design and Implementation of Chess-Playing Robotic System", IJCSET(www.ijcset.net) | May 2015 | Vol 5, Issue 5,90-98.

[10] Gurjit Kaur, Anil K. Yadav, VarunAnand, "Design and Implementation of Artificially Intelligent Microcontroller based Chess Opponent", Proceedings of the World Congress on Engineering 2010 Vol I WCE 2010, June 30 - July 2, 2010, London, U.K.

[11] Guofeng Tong, Ying Qu, Tong Cheng, "Human-computer interactive gaming system - a Chinese chess robot" Intelligent Robots and Systems, 2009, IEEE/RSJ International Conference.

[12] SitiZarinaMohdMuji, MohdHelmy Abdul Wahab, RadziAmbar, "Design and Implementation of Electronic Chess Set", International Conference on Advances in Electrical, Electronic and System Engineering, 14-16 Nov 2016, Putrajaya, Malaysia, 2016.

[13] Zhang Baofeng, Wang Ya, Zhu Junchao "Design of High Speed Data Acquisition System Based on FPGA and DSP" IEEE 2010.

[14] Jing pang "intelligent traffic light controller design using PFGA digest of technical papers", international Conference on consumer electronics (ICCE), IEEE, 2016.

[15] J. Zurada, "Introduction to Artificial Neural Network System", West Publishing Company, ISBN: ISBN 0-3 14-93391 -3, 1992.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 195-203 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Guidance Law Based on Line-of-Sight Rate Information Considering Uncertain Modeled Dynamics

Saori Nakagawa*, Takeshi Yamasaki, Hiroyuki Takano, Isao Yamaguchi

National Defense Academy of Japan, Department of Aerospace Engineering, 239-8686, Japan

ARTICLEINFO

Article history:
Received: 28 August, 2018
Accepted: 28 October, 2018
Online: 15 November, 2018

Keywords:
Aerospace Engineering
Autonomous system
Adaptive control and tuning

ABSTRACT

Proportional navigation (PN) is a widely-used guidance law for missile-target engagement. The goal of the missile intercept problem is to reduce the closest distance between the missile and target by diminishing the line-of-sight rate (LOS rate). In general, PN guidance law necessitates information of the LOS rate and missile velocity. The closing velocity (relative approaching speed to the target) instead of the missile velocity is an additional option for effective guidance. However, there are cases where a sensing device for measuring target motions that can be mounted on a missile is limited. In this paper, we propose a novel guidance law on the basis of proportional navigation (PN) using only line-of-sight (LOS) rate information. In this paper, an uncertainty and disturbance estimator (UDE) is applied to estimate such target motions including velocity change or unpredictable movement etc. The UDE works also for compensating uncertain modeled dynamics such as a missile's bearing uncertainty and velocity changes. The proposed guidance law is referred to as uncertainty and disturbance-compensated intercept guidance. Numerical simulations with some engagement scenarios are presented taking account of the velocity changes of the missile to demonstrate the potential of the proposed guidance law.

1 Introduction

Proportional navigation (PN) is a widely-used guidance law for terminal interception of target-missile engagement [1], and the guidance law attracts attention since the first half of the 20th century. Researches have been conducted not only for missile guidance, but also in the fields such as the guidance of vehicles [2], the formation flight of aircraft [3], small UAVs autonomous path-following [4], furthermore the problem of rendezvous of satellites and terminal guidance control to a small moon landing [5, 6].

Proportional navigation (PN) is mainly classified into two; (1) true proportional navigation (TPN) using closing velocity information and (2) the pure proportional navigation (PPN) using the missile's velocity information [7]. Both laws use information on the line-of- sight (LOS) rate (rate of change in the direction of a line connecting a missile and a target) and the velocity of the missile (if it is obtained, using the closing velocity to the target). The PN is widely used as a guidance law for the intercept in two-vehicle engagement because the implementation only with the LOS rate information is quite simple [1]. In the case of a stationary or non-maneuvering target, without measurement error, dynamics lag, nor lateral acceleration limit on the navigation system of the missile, the PN can completely intercept the target with zero miss distance. For such a reason, researches on missile guidance have been conducted based on the PN or alternative representation such as the zero-effort-miss (ZEM) guidance. For instance, the augmented proportional navigation (APN) compensates a target lateral constant acceleration by adding a modified acceleration term to the PN law [8]. The compensated PN [9] (or velocity change-compensated PN [10]) guides missiles to keep the LOS rate constant at zero by correcting the LOS rate change produced by the missile's axial acceleration.

In [11], optimal theory on intercept guidance provides that the effective navigation constant of the PN set to three is an optimal against the non-maneuvering target in the sense of the least squared integral of the lateral accelerations. The preceding guidance system does not take dynamics-lag into account while an opti-

*Saori Nakagawa, 1-10-20 Hashirimizu, Yokosuka, Kanagawa 239-8686, Japan, TEL:+81-468-41-3810, email: saori.sari.nkgw@gmail.com

https://dx.doi.org/10.25046/aj030626

mal guidance for a first-order time-lag system is given in [8]. The guidance law in [8] that minimizes the squared integral of the commanded lateral acceleration of the missile uses the target lateral acceleration information. In addition, in this optimal guidance law (referred to as OG in this paper), a variable gain is used instead of the constant value, and the variable gain can be expressed as a function of the time-to-go (the time remaining till interception). In this regard, the suboptimal guidance (SOG) law was proposed using PN with phase lead compensation by Baba et al. [12]. The optimal or suboptimal guidance laws described so far are effective under the assumption where the closing velocity information is available. Sensing device for target information may depends on the missile's operations. There are cases where the missile cannot install an equipment measuring the closing velocity. Therefore, in this study, we focus on the case where only LOS rate can be acquired as for target maneuvers-related information other than that of the missile itself.

Since the closing velocity cannot be obtained while only LOS rate is used, the usual fixed gain approach in PN may not exhibit sufficient guidance performance. Here, we focus on the LOS rate dynamics in order to enhance or keep the guidance performance under restricted conditions where only information as for the target is to be the LOS rate. We proposed a new guidance law that drives the LOS rate to be zero, estimates and compensates an uncertain group including unmodeled dynamics and external [13]; In [13], the authors did not consider time-lag in system response, so we present results for a time-lag system [14]. Furthermore, this study demonstrates the robustness of the proposed guidance affected by the axial acceleration by the missile thrust and the deceleration due to aerodynamic drag.

The rest of this paper is organized as follows: Section 2 details the setting of the guidance problem with notation assignments for the missile-target engagement, followed by the governing equations. Next, a guidance law, which is based on an assembled simple LOS rate dynamics model, is developed. In Section 3, numerical simulation results are shown to demonstrate the potential of the proposed guidance law. Section 4 summarizes this study.

2 Guidance Theory

2.1 Engagement problems and Governing equations

This section introduces the notations, assumptions, governing equations for the guidance law, and defines the engagement problem.

Figure 1 shows a missile-target geometry and their related notations for deriving the guidance law for the missile-target engagement scenario. Here after, the guided vehicle is referred to as a missile. λ denotes the LOS angle ,and R is the LOS range. V, γ and a for each vehicle represents the velocity, flight path angle, and lateral acceleration, respectively. The subscripts for these variables distinguish the missile from the target; with "m" for the missile, and "t" for the target.

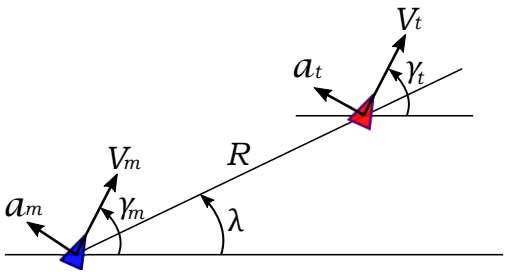


Figure 1: Engagement Geometry

Assumptions are made for the derivation of the proposed guidance law as

(i) The missile and target are point masses, moving in a plane.

(ii) The target moves with a constant speed.

(iii) The LOS rate is available without error.

(iv) The response lag to the commanded lateral acceleration for the missile can be represented by a first-order-lag system.

Under these assumptions, the governing equations on the missile-target engagement can be summarized as

$$\dot{R} = V_t \cos(\gamma_t - \lambda) - V_m \cos(\gamma_m - \lambda) \tag{1}$$

$$\dot{\lambda} = \frac{1}{R}\{V_t \sin(\gamma_t - \lambda) - V_m \sin(\gamma_m - \lambda)\} \tag{2}$$

$$\dot{\gamma}_t = a_t / V_t \tag{3}$$

$$\dot{\gamma}_m = a_m / V_m \tag{4}$$

$$\dot{a}_m = -\frac{1}{\tau_1} a_m + \frac{1}{\tau_1} a_c \tag{5}$$

$$\dot{V}_m = \frac{T - D}{m} \tag{6}$$

$$\dot{m} = \frac{-T}{g I_{sp}} \tag{7}$$

where a_c is a commanded value of missile's lateral acceleration, and system response a_m is approximated by the output from a first-order-lag system with the time constant τ_1. Also T, D, I_{sp} and g denote the missile's thrust, drag force, and specific impulse, and the gravitational acceleration, respectively. For additional variables for simulations, the lift denoted by L and the drag-related equations are shown as;

$$L = \frac{1}{2}\rho {V_m}^2 S C_L \tag{8}$$

$$D = \frac{1}{2}\rho {V_m}^2 S C_D \tag{9}$$

$$C_D = C_{D_0} + \kappa {C_L}^2 \tag{10}$$

where ρ is the air density, S is the cross sectional area, C_L, C_D, C_{D_0}, and κ denote the lift coefficient, the drag coefficient, the zero-lift drag coefficient, and the induced drag-related parameter, respectively.

If the LOS angle λ is maintained constant, the missile can intercept the target [1, 8]. That is, the problem for deriving the guidance law on missile-target engagement is to calculate the commanded lateral acceleration for missile, a_c, which will bring the LOS angular velocity (the LOS rate) $\dot{\lambda}$ to zero. Thus, the objective of this study is to find a_c nullifying the LOS rate without using information of the closing velocity.

2.2 Construction of a simplified dynamics model

In designing a guidance system, the construction of a dynamics model plays an important role. An accurate model is indispensable to build a high precision guidance. On the other hand, a designed guidance system with such an accurate model tends to become a complicated and high-order system, and to demand multiple sensors and/or estimators. For this reason, we try to use a simplified guidance model as much as possible aiming at simplifying the design process for a guidance system that uses only information of the LOS rate.

In order to design a guidance system, this study focuses on a mathematical model representing the LOS rate dynamics. By differentiating Eq. (2) and substituting Eqs. (1), (3) and (4), the LOS rate dynamics model is derived in the following differential equation form;

$$\ddot{\lambda} = \frac{1}{R}\left\{-2\dot{R}\dot{\lambda} - a_m + h_0\right\} \tag{11}$$

$$\begin{aligned} h_0 = {} & a_m\{1-\cos(\gamma_m-\lambda)\} + a_t\cos(\gamma_t-\lambda) \\ & + \dot{V}_t\sin(\gamma_t-\lambda) - \dot{V}_m\sin(\gamma_m-\lambda) \end{aligned} \tag{12}$$

The term h_0 in the preceding Eq. (12) includes "model uncertainty" arising from mathematical simplifications, and "disturbance" caused mainly from uncertain target accelerations or maneuvers. Hereafter, the term h_0 including such model uncertainty and external disturbance is called an uncertainty and disturbance term (UDT).

In [15], Eq. (11) is used as a system equation for the development of the guidance logic. However, Eq. (11) includes not only the available LOS rate but also the LOS range and the closing velocity ($V_c = -\dot{R}$) that cannot be measured or estimated. Instead of Eq. (11), in order to take account of the restriction where only the LOS rate is available, this study uses the following equation;

$$\ddot{\lambda} = -\frac{a_m}{R_0} + h \tag{13}$$

where R_0 is a given parameter in advance, for example the reachable flight range of the missile or the initial value of the LOS range. The UDT (h_0) of Eq. (11) and additional uncertain terms caused from this simplification are combined as

$$h = \left(\frac{1}{R_0} - \frac{1}{R}\right)a_m + \frac{1}{R}\left\{-2\dot{R}\dot{\lambda} + h_0\right\} \tag{14}$$

The estimation of the UDT (h) is explained in Section 2.4.

In this study, the command lateral acceleration a_c is considered as the input to the guidance system while taking account of the missile dynamics which is approximated as a first-order-lag system. In the guidance system, the LOS rate $\dot{\lambda}$ and the lateral acceleration of the missile a_m are treated as state variables. Thus, the lateral acceleration a_m is treated as an intermediate variable for a backstepping approach. The purpose of this guidance system can be interpreted as "driving the first state variable $\dot{\lambda}$ to zero with the input of a_c". A method for calculating the input a_c driving the first state variable to zero will be described in the next section.

2.3 Derivation of uncertainty and disturbance-compensated guidance

In this section, we explain the derivation of uncertainty and disturbance-compensated guidance based on the governing Eqs. (5) and (13). The guidance law should drive the LOS rate $\dot{\lambda}$ to zero with the lateral commanded acceleration a_c as described in the previous section. However, the first state variable cannot be directly controlled by the lateral acceleration commanded value a_c. Therefore, this study uses the backstepping method to control the LOS rate $\dot{\lambda}$ via the intermediate state a_m. A candidate of the Lyapunov function is defined as

$$V = \frac{1}{2}\sigma^2 \tag{15}$$

where, for simplicity, the LOS rate $\dot{\lambda}$ is defined as σ. When Eq. (15) is differentiated with time, it becomes

$$\dot{V} = \sigma\dot{\sigma} = \sigma\left(-\frac{a_m}{R_0} + h\right) \tag{16}$$

The estimated value of the UDT h is defined as $\hat{h}$. Moreover, the desired missile lateral acceleration a_{md} is set as

$$a_{md} = R_0\left(k\sigma + \hat{h}\right) \tag{17}$$

On the other hand, addition and subtraction of a_{md}/R_0 to Eq. (13) and making $(a_{md} - a_m)$ term lead to

$$\dot{\sigma} = -\frac{a_{md}}{R_0} + h + \frac{a_{md} - a_m}{R_0} \tag{18}$$

Substituting Eq. (17) for the first term of Eq. (18) provides

$$\dot{\sigma} = -k\sigma + \left(h - \hat{h}\right) + \frac{a_{md} - a_m}{R_0} \tag{19}$$

Substituting Eq. (19) for the first equality in Eq. (16), it is found that

$$\dot{V} = -k\sigma^2 + \sigma\left(h - \hat{h}\right) + \frac{\sigma}{R_0}\left(a_{md} - a_m\right) \tag{20}$$

If the lateral acceleration a_m converges to the desired lateral acceleration a_{md}, and when the UDT(h) can be estimated without error, that is, $\hat{h}\rightarrow h$, the Eq. (20) is expressed as $\dot{V} \approx -k\sigma^2 \leq 0$ (the equality holds when $\sigma = 0$). Consequently, if k is positive, V can be called as the Lyapunov function. For this reason, the LOS rate σ converges to zero as the lapse of time.

In order to make a_m close to a_{md}, a new candidate Lyapunov function is defined as

$$V_1 = \frac{1}{2}\sigma^2 + \frac{1}{2}(a_{md} - a_m)^2 \tag{21}$$

Differentiating Eq. (21) with time leads to

$$\dot{V}_1 = \sigma\dot{\sigma} + (a_{md} - a_m)(\dot{a}_{md} - \dot{a}_m) \tag{22}$$

From the observation of Eqs. (5), (17) and (22),commanded lateral acceleration can be selected as

$$a_c = a_{md} + \frac{\tau_1\sigma}{R_0} \tag{23}$$

If it can be approximated as $\dot{a}_{md} \fallingdotseq 0$ and the estimated value become $\hat{h}\rightarrow h$, Eq. (22) becomes

$$\dot{V}_1 = -k\sigma^2 - (a_{md} - a_m)^2/\tau_1 \tag{24}$$

that shows $\dot{V}_1 \leq 0$.

Equations (17) and (23) represent the guidance law. Excepting the estimated value $\hat{h}$ of the UDT in Eqs. (17) and (23), the guidance law includes terms proportional to the LOS rate without the closing velocity: Eqs. (17) and (23) essentially includes the PN guidance law.

For comparison purpose to the proposed guidance law, the conventional guidance laws are illustrated next, which clarifies what kind of extra information is required for each guidance law.

The PN guidance law that uses missile velocity instead of the closing velocity (what is called 'pure PN (PPN)') is represented as [16]:

$$a_{mPPN} = NV_m\sigma \tag{25}$$

where N denote the navigation constant. The lateral acceleration is assumed to be generated perpendicular to the velocity vector in the PPN guidance law. In this research, the closing velocity information cannot be obtained. However, if this is obtained, it is effective to use the closing velocity V_c instead of the missile velocity V_m of the Eq. (25). Thus, the command lateral acceleration of PN guidance law is

$$a_{mTPN} = N'V_c\sigma \tag{26}$$

where N' is the effective navigation constant [16]. According to the optimal control theory as described in the introduction [11], if the missile has perfect dynamics (it has no response lag), and the target has constant-speed with rectilinear motion, when the effective navigation constant is 3 ($N' = 3$), the input of command lateral acceleration is minimizing the square integral of acceleration.

The augmented PN (APN) guidance law is a well-known optimal guidance law against the constant maneuvering target. The target's constant lateral acceleration-related term is added to the TPN guidance law as [8]

$$a_{mAPN} = N'V_c\sigma + \frac{N'a_t}{2} \tag{27}$$

where a_T indicates the target maneuver (lateral acceleration which is assumed to be normal to the LOS) [15].

Furthermore, the optimal guidance law (OG) is derived based on the optimal control theory for a liner system in case where the response of the system can be approximated by a first-order-lag;

$$a_{mOG} = \frac{N'}{t_{go}^2}[y + \dot{y}t_{go} + 0.5a_t t_{go}^2 - a_c{\tau_1}^2(e^{-x} + x - 1)] \tag{28}$$

where

$$x = \frac{t_{go}}{\tau_1} \tag{29}$$

Also, y indicates a component perpendicular to the initial LOS. y is subtraction of the missile position component from the target one with respect to the initial LOS. If y is sufficiently small, the relationship of trigonometric ratio provides

$$\lambda \fallingdotseq \sin\lambda = \frac{y}{R} \tag{30}$$

The following relationship can be obtained by differentiating both sides of Eq. (30).

$$V_c\dot{\lambda} \fallingdotseq \frac{y + \dot{y}t_{go}}{t_{go}^2} \tag{31}$$

The numerator in the right hand side in Eq. (31) is called ZEM (the miss distance with no guidance force applied). The variable gain in Eq. (28) is derived as

$$N' = \frac{6x^2(e^{-x} - 1 + x)}{2x^3 + 3 + 6x - 6x^2 - 12xe^{-x} - 3e^{-2x}} \tag{32}$$

Using Eq. (31), Eq. (28) is expressed as

$$a_{mOG} = N'V_c\sigma + \frac{N'a_t}{2} - N'a_c x^{-2}(e^{-x} + x - 1) \tag{33}$$

In the simulation of this paper, these commanded acceleration (Eqs. (25) ~ (27) and Eq. (33)) are used to compare with the proposed guidance law.

k of the proposed guidance law (Eq. (17)) is a positive constant to be tuned. Some design policy can be helpful. In this study, the gain is determined as compared with PN guidance law of Eq. (25) as

$$k = NV_{m0}/R_0 \tag{34}$$

where R_0 is a pre-specified value; it could be the initial LOS range or the possible flight range. V_{m0} and N denote the assumed average velocity of the missile and

the navigation constant, respectively. In this proposed guidance law, N is the main tuning parameter.

2.4 Estimation of the uncertainty and disturbance term

We describe the estimation method [15, 17, 18] for UDT h in this section.

As for the UDT estimation, assembled term for uncertainty and disturbance in a focused system dynamics are estimated based on the system dynamics. For example, a discrete-type method, so-called time delay control (TDC) [19, 20] was proposed for an uncertainty and disturbance compensator. The TDC technique, however, arose a potential problem; since the time derivative in TDC is approximated with numerical differentiations, instability of the system may occur. In [17], a frequency-domain-based approach that can be treated in continuous system were proposed by Zhong. In this method, the stability of the disturbance is guaranteed. Hence, this study follows the Zhong's idea for the UDT estimator with additional extensions for the measurement limit.

In [18] and [15], the authors also proposed missile guidance law using an uncertainty and disturbance compensator under the condition where the measurements of the LOS range and the range rate as well as the LOS rate can be obtained. The guidance system proposed here imposed further limitations, that is, the proposed system can use information of the LOS rate only.

The UDT estimation methodology using the LOS rate without the LOS range nor the closing velocity is developed.

The first step begins with solving Eq. (7) for h

$$h = \dot{\sigma} + \frac{a_m}{R_0} \tag{35}$$

The right hand side of the Eq. (35) contains the differential term of the LOS rate which is a measured variable. Using the derivative of the measured variable (or estimated value) may cause numerical problem since the differentiation is sensitive to measurement noise and to the characteristics of state estimator. Thus, using the derivative of the LOS rate should be avoided.

Similar to [18] and [15], assuming that a UDT h is the input to some strictly proper virtual filter $G(s)$, and setting its output denoted by $\hat{h}$ as the estimate of h, lead to the following representation in the Laplace domain form as

$$\hat{H}(s) = G(s)H(s) \tag{36}$$

In fact, it is impossible to extract $h(t)$ as a signal. In this paper, we merely assume that it is the input to the virtual filter for the sake of developing the UDT estimator. Although various kinds of filters can be the candidates for the virtual filter, the following first-order-lag filter is applied in order to facilitate the design.

$$G(s) = \frac{1}{\tau s + 1} \tag{37}$$

where τ indicates the time constant and that is a design parameter to be tuned. The time constant τ characterizes the convergence rate. By using the first-order-lag filter shown in Eq. (37), Eq. (36) is expressed in the time domain form as follows:

$$\tau\dot{\hat{h}}(t) + \hat{h}(t) = h(t) \tag{38}$$

where we assume $\hat{h}(0) = 0$ without loss of generality. Substitute the left hand side of Eq. (38) for the left hand side of Eq. (35) and rearrange it to derive

$$\tau\dot{\hat{h}} + \hat{h} = \dot{\sigma} + \frac{a_m}{R_0}$$

$$\tau\dot{\hat{h}} = \dot{\sigma} - \hat{h} + \frac{a_m}{R_0} \tag{39}$$

After integration of Eq. (39) with intervals $[0, t]$ and solving for $\hat{h}$, the following estimator can be obtained.

$$\hat{h} = (\sigma - \sigma(0) + \omega)/\tau \tag{40}$$

$$\dot{\omega} = -\hat{h} + \frac{a_m}{R_0}, \omega(0) = 0, \hat{h}(0) = 0 \tag{41}$$

The estimator, denoted by Eqs. (40) and (41) is consisted only of the measurable LOS rate, the lateral acceleration of missile and the design parameter. Therefore, if the LOS rate is measurable, the UDT estimator is designed using one design parameter τ.

The convergence rate of the UDT can be enhanced by decreasing the design parameter τ. However, in practically, it has minimum limitation to attenuate the influence of noise. Since Zhong [17] explains the convergence property of the estimate, it is omitted in this paper.

2.5 Stability of the Uncertainty and Disturbance Term

In this section, we explain about the convergence of the uncertainty and disturbance estimator. It is assumed that the UDT (h) is continuous (the time differentiation of h is bounded). Estimated error is defined as

$$\tilde{h} \equiv h - \hat{h} \tag{42}$$

Considering the following non-negative-definite function.

$$V_h = \frac{\tau}{2}\tilde{h}^2 \tag{43}$$

and performing the time derivative of Eq. (43) yields

$$\dot{V}_h = \tau\tilde{h}\dot{\tilde{h}} \tag{44}$$

Substituting Eq. (13) into Eq. (39) provides

$$\dot{\tilde{h}} = \frac{1}{\tau}\tilde{h} \tag{45}$$

Then, the differentiation with time of Eq. (42) derives

$$\dot{\tilde{h}} = \dot{h} - \dot{\hat{h}} \tag{46}$$

Substituting Eq. (46) for Eq. (45) leads to

$$\dot{\tilde{h}} = \dot{h} - \frac{1}{\tau}\tilde{h} \tag{47}$$

Noting Eq.(47), Eq. (44) can be rewritten as

$$\dot{V}_h = \tilde{h}\left(\tau\dot{h} - \tilde{h}\right) \tag{48}$$

where $\dot{h}$ is assumed to be bounded, considering $|\dot{h}| \le C$ (C: the positive constant), $\dot{V}_h < 0$ is guaranteed when $|\tilde{h}| > \tau|\dot{h}|$ and the estimation error with elapse of time ($t \to \infty$) is ultimately bounded within the range of

$$|\tilde{h}| \le \tau|\dot{h}| \le \tau C \tag{49}$$

It can be seen from Eq. (49) that the estimation error can be reduced by decreasing τ. Furthermore, when the UDT is constant, that is $\dot{h} = 0$, the uncertainty and disturbance term h is estimated without error ($\hat{h} \to h$).

2.6 Discussion of uncertainty and disturbance-compensated intercept guidance with LOS rate measurements

From the UDT estimator of Eqs. (40) ~ (41) and the guidance command of Eq. (35), the proposed UDT is summarized as follows:

$$\dot{\omega} = -\hat{h} + \frac{a_m}{R_0}, \omega(0) = 0, \hat{h}(0) = 0$$
$$\hat{h} = (\sigma - \sigma(0) + \omega)/\tau$$
$$a_m = R_0(k\sigma + \hat{h})$$

When the lateral acceleration command value is not limited and there is no time-lag in the guidance system, the lateral acceleration in Eq. (41) equals to the desired lateral acceleration of Eq. (17) as described in the preceding equation. In this case, Eq. (17) is substituted into Eq. (41) with $a_m = a_{md}$ then one obtain

$$\dot{\omega} = k\sigma, \omega(0) = 0 \tag{50}$$

As shown above, the UDT estimator proposed by the authors is almost equivalent to the one composed only proportional-integral filter of the LOS rate, and the design parameters k, R_0 and τ. That is, in comparison with the existing guidance laws demanding extra information such as the closing velocity, the LOS range, the time-to-go and the lateral acceleration of target, the proposed guidance law using only the LOS rate is useful from the view point of implementation simplicity.

3 Simulation

Guidance simulations have been performed to confirm the effectiveness of the proposed guidance law. When the missile velocity is sufficiently larger than the target velocity, there is no significant difference between the proposed guidance law and PN guidance law. Therefore, this paper presents an example case; the target velocity is larger than the missile velocity where the performance difference appears noticeably by using the proposed guidance law. In the simulations, the equations of motion Eqs. (1) ~ (7) are used. The setting values of initial states and design parameters are summarized in the Table 1.

The limitation load of the missile is set to forty times of the gravitational acceleration (G), that is, 40G. When the commanded lateral acceleration exceeds the limit, the limited value is used as a command in the simulation. The target acceleration set as 4G to be constant. For comparison purposes, the PPN guidance law using the missile velocity in Eq. (25); the TPN guidance law using the closing velocity in Eq. (26); the APN guidance law adding the extra term which considered the maneuvering target in Eq. (27); and the OG of using t_{go} information in Eq. (33) are compared by the proposed guidance law in the simulations.

Table 1: Initial kinematics and design parameters

Parameters	Values	
V_m	1,500	[m/s]
x_m	0	[m]
y_m	0	[m]
γ_m	20	[deg]
$\max\{a_m\}$	40G	[m/s^2]
m	100	[kg]
$\min\{m\}$	60	[kg]
S	0.01267	[m^2]
I_{sp}	300	[s]
G	9.806	[m/s^2]
V_t	3,500	[m/s]
x_t	55,000	[m]
y_t	9,000	[m]
γ_t	-175	[deg]
a_t	4G	[m/s^2]
k	NV_{m0}/R_0	[1/s]
N	3	[−]
N' (TPN, APN)	3	[−]
τ	0.3	[s]
τ_1	0.5	[s]
R_0	$R(0)$	[m]

The simulations with such comparative guidance laws are made because the simulation setting of guidance laws with the effective navigation constant of $N' = 3$ in TPN and APN provide optimal solution in the sense that the square integral of lateral acceleration becomes minimum when there is no system response delay. The proposed guidance law assumes that only LOS rate information is used, whereas the TPN, APN and OG guidance laws use the closing velocity that cannot be used in this assumption. Hence, simulation results PPN using the missile velocity are also

shown as for another comparison guidance law with the same measurement condition. When simulations are performed for TPN, APN and OG guidance law, information on the closing velocity and the target lateral acceleration can be used without error.

Table 2: Resulting miss distance is shown. For comparison, we also show the results when V_m is constant, that is, when the affect of the axial acceleration by the thrust or of the deceleration due to the aerodynamic drag is not considered.

guidance method	miss distance [m]	
	V_m=const.	V_m changes
Proposed	$< \delta$	$< \delta$
PPN	565.76	813.19
TPN	0.18	8.43
APN	0.0028	0.02
OG	$< \delta$	0.0096

$\delta := 0.001$

The flight trajectories of the missile and the target are shown in Fig.2. The red solid line drawn from the upper right to the lower left is the trajectory of the target, the red broken line denotes the missile trajectory with the PPN guidance law, the green broken dotted line is made with the TPN guidance law, the magenta dotted line is made with the APN guidance law, the red broken line is made with the OG law, and the blue solid line is made with the proposed guidance law. The resulting miss distance value using each guidance law is shown in Table 2. For comparison purpose, the results of the case where the missile velocity is constant is also shown. In this case, when the miss distance is 0.001[m] or less, it is expressed as "$< \delta$".

From enlarged view of the lower right of Fig. 2 (the equivalent aspect ratio), it is observed that the trajectories with the APN, OG and proposed guidance laws have similar trajectories. Simultaneously, we can see that the APN, OG and the proposed guidance law approach to the target in a shape close to a straight line, but the PPN and TPN guidance law draws a curve towards the target. It is considered that this causes large miss distance values as shown in Table 2. Additionally, for the case where the missile velocity changes with time, the miss distance values of the PPN, TPN, APN and OG guidance law become large, whereas the proposed guidance law keeps in the same level. Note that the trajectory with the proposed guidance law is almost similar to the trajectories with the APN guidance law and OG law using the closing velocity and the lateral acceleration, and the miss distance value is smaller than the other guidance laws.

The curvature of the missile trajectory in Fig. 2 becomes clear by observing the histories of the lateral acceleration applied to the missile. Figure 3 shows the time history of the missile's lateral acceleration (so-called latex) when using the each guidance law as well as the time history of the LOS rate in Fig. 4. In Fig. 3, the value of lateral accelerations of the APN guidance and OG laws converge to almost zero with exception at the time of intercept. In the proposed guidance law, since the target lateral acceleration is considered as a part of the UDT, it turned out that the missile lateral acceleration tends to converge to a value close to the target lateral acceleration at 4G. In addition, it can be seen from Fig. 3 that the PPN and TPN guidance laws have reached the limited value of 40G before the intercept. The PPN guidance law is considered to generate a large miss distance value as shown in Table 2 because it reached this limit earlier than that of the TPN guidance law. It can also be observed from Fig. 4, since the LOS rate with the proposed guidance law converges to a value close to zero at the impact time, a small miss distance value shown in Table 2 was obtained. Consequently, under the condition shown in Table 1, that is, when the target velocity is larger than the missile velocity, the proposed guidance law is considered to exhibit desirable performance.

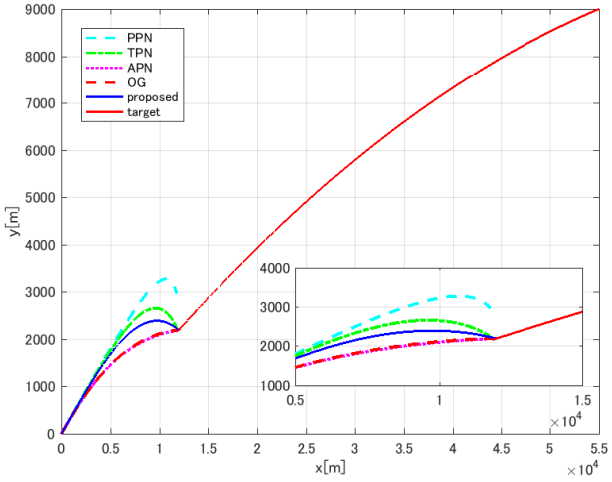


Figure 2: Flight trajectories of the missile and the target. The lower right is an enlarged view.

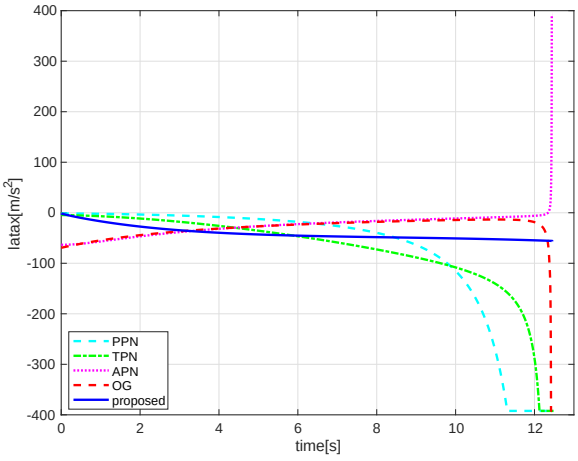


Figure 3: The time history of the missile's lateral acceleration in each guidance law is shown. The limitation load factor of the missile is set 40G this is forty times of the gravitational acceleration.

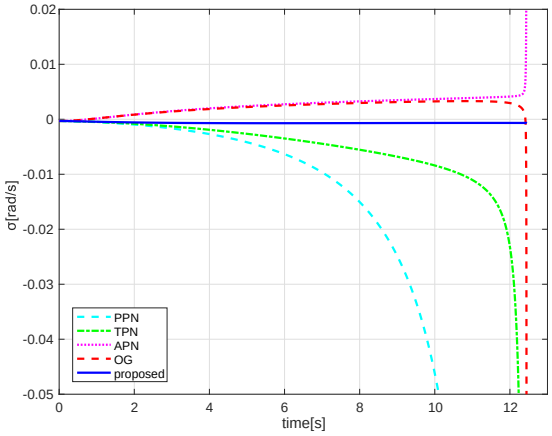


Figure 4: LOS rate vs time is shown. It is desirable for the LOS rate to be zero.

Figure 5 and 6 shows the time histories of the missile velocity and the missile mass. In this simulation, since the minimum value of the missile mass is limited to 60[kg], the missile intercepted the target before reaching that value.

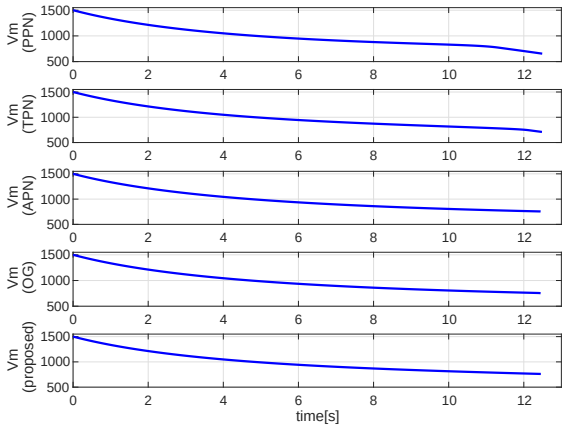


Figure 5: The missile velocity of each guidance considering axial acceleration by the thrust force, and deceleration due to aerodynamic drag vs time.

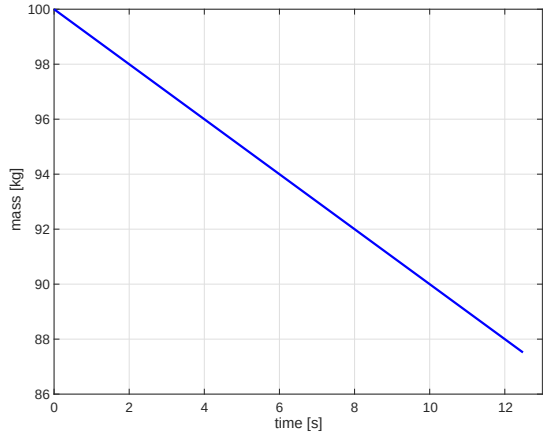


Figure 6: Mass of missile decreasing with time.

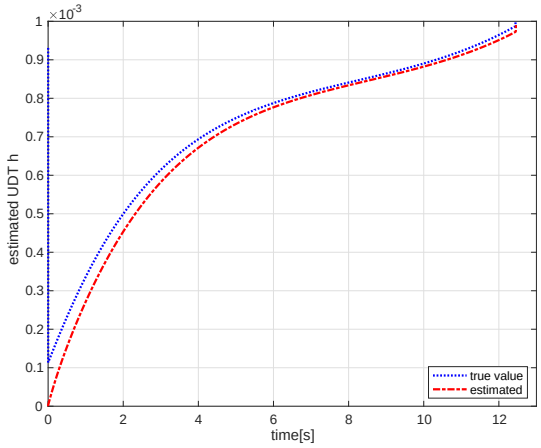


Figure 7: The time history of the estimated UDT h and the true value.

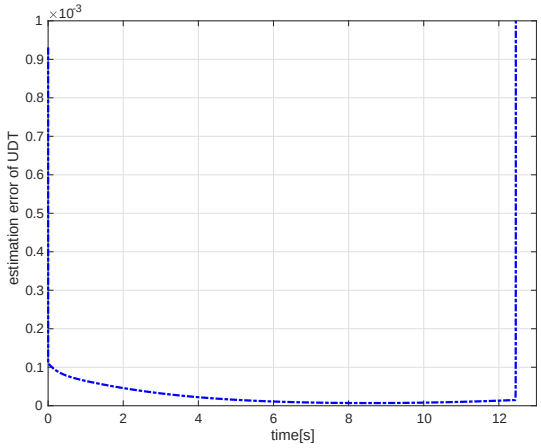


Figure 8: UDT error vs time. This figure shows $(true) - (estimate)$ values in Fig.7.

Finally, to demonstrate the performance of the UDT estimator, the time history of the UDT h which is a true value and that of estimated value $\hat{h}$ are shown in Fig. 7, in addition, the time history of the estimation error is also shown in Fig. 8. From these figures, it can be seen that the estimator using the time constant of $\tau = 0.3$[s] can be estimated so that the estimated value approaches the true value about 2 to 6 seconds. By reducing the value of the time constant τ, it can be expected to increase the convergence rate to the true value, but the response of the lateral acceleration tends to fluctuate. Therefore, the time constant is too small is not desirable. The reason that the estimation error increases at the intercept time in Figs. 7 and 8 is that numerical divergence occurs for the real system where the distance R in Eq. (14) converges to zero while the acceleration converges to the target acceleration that are none zero value. Therefore, the proposed guidance law should be deactivated or hebetated at the end of the interception to obtain better performance. This is the future work for this study since the LOS range is not measured. Regardless of the difficulties, the result-

ing miss distance as shown in Table 1 demonstrates the excellent performance of the proposed guidance law.

4 Conclusions

A new guidance law for missile-target engagement using information of LOS rate only is developed. From the simulation study, the proposed guidance law exhibits excellent performance almost equal to the OG law where the OG law uses additional information of the LOS range, the closing velocity and the lateral acceleration of the target, whereas the proposed guidance law uses only LOS rate information even if uncertain modeled dynamics such as velocity change and directional error is included. Hence, even with a guidance system that cannot measure the closing velocity or maneuvering target, we can anticipate that almost the same performance as that of conventional excellent guidance with using the closing velocity or target maneuver can be obtained by using the proposed guidance law.

For future works, we will apply the proposed guidance against various targets or initial settings, study a blind sight determination method (or deactivation technique in the proximity of the target), and investigate the influence of the time-lag constant changes and observation noise to the design parameters on guidance performance.

References

[1] R.E. Machol, W.P. Tanner.Jr, and S.N. Alexander, System Engineering Handbook, Chap.19 Guidance, R.E. Hill, 1965.

[2] Y. Yavin, R.De. Villiers, Proportional navigation and the game of two cars: the case of a pursuer with variable speed, Computers & Mathematics with Applications, Vol.18, No.1-3, pp.69-76, 1986.

[3] M.J. Tahk, C.S. Park, and C.K. Ryoo , Line-of-Sight Guidance Laws for Formation Flight, Journal of Guidance, Control, and Dynamics, Vol.28, No.4, pp.708-716, 2005.

[4] Y. Sato, T. Yamasaki, et al., Trajectory Guidance and Control for a Small UAV, International Journal of Aeronautical and Space Sciences, Vol.7, No.2, pp.137-144, 2006.

[5] M. Matsuda, H. Takano, T. Yamasaki and I. Yamaguchi, A Note on the Optimal Spacecraft Guidance, APISAT, 7.4.1.pdf, 2012.

[6] P.J. Yuan and S.C. Hsu, Rendezvous Guidance with Proportional Navigation, Journal of Guidance, Control, and Dynamics, Vol.17, No.2Cpp.409-411, 1994.

[7] C.D. Yang, C.C. Yang, A Unified Approach to Proportional Navigation, Transactions on aerospace and electronic system, IEEE, Vol.33, No.2, pp.557-567, April 1997.

[8] P. Zarchan, Tactical and Strategic Missile Guidance, Sixth Edition, AIAA, ch.8, pp.163-185, 2007.

[9] P. Zipfel, Modeling and Simulation of Aerospace Vehicle Dynamics, Second Edition, AIAA, 2007.

[10] T. Kuroda, and F. Imado, Advanced Missile Guidance System against a Very High Speed Maneuvering Target, Proceedings of AIAA Guidance, Navigation and Control Conference, pp.3445, 1989.

[11] A.E. Bryson, Jr., and Y.C. Ho, Applied Optimal Control, - Optimization, Estimation, and Control-, Hemisphere Publishing Corporation, New York, pp.154-155, 1975.

[12] Y. Baba, K. Inoue, and R.M. HOWE, Suboptimal guidance with line-of- sight rate only measurements, Proceedings of AIAA Guidance, Navigation and Control Conference, AIAA 88-4066-CP, Minneapolis, pp.122-129, 1988.

[13] S. Nakagawa, T. Yamasaki, H. Takano, and I. Yamaguchi, Disturbance- Compensated Intercept Guidance Using Line-of-Sight Rate Information, 11th Asian Control Conference, IEEE, pp.388-393, December 2017.

[14] S. Nakagawa, T. Yamasaki, H. Takano, and I. Yamaguchi, Comparison study on Disturbance-Compensated Intercept Guidance Using Line-of-Sight Rate Information against Optimal Guidance, Multi-Symposium on Control System, 2018.

[15] T. Yamasaki, S.N. Balakrishnan, and H. Takano, Sliding mode-based intercept guidance with uncertainty and disturbance compensation, Journal of Franklin Institute, Vol. 352(11), pp. 5145-5172, 2015.

[16] C.D. Yang, and C.C. Yang, A Unified Approach to Proportional Navigation, IEEE , Transactions on aerospace and electronic system, Vol.33, No.2, pp.557-567, April 1997.

[17] Q.C. Zhong, Comments on gA time delay controller for systems with Uncertain Dynamicsh, 27th January 2013.

[18] S.B. Phadke, and S.E. Talole, Sliding Mode and Inertial Delay Control Based Missile Guidance, IEEE Transactions on Aerospace and Electronic Systems, Vol.48(4), pp.3331-3346, 2012.

[19] K. Youcef-Toumi, and O. Ito, Controller design for systems with unknown nonlinear dynamics, Proceedings of the American Control Conference, Minneapolis, pp.836-845, 1987.

[20] K. Youcef-Toumi, and O. Ito, A time delay controller for systems with unknown dynamics, Journal of Dynamic System, Measurement, and Control, - Transactions of the ASME, Vol.112(1), pp.133-142, 1990.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 204-212 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Attitude Control Simulation of a Legged Aerial Vehicle Using the Leg Motions

Yoshiyuki Higashi[*,1], Soonki Chang[2]

[1]*Kyoto Institute of Technology, Faculty of Mechanical Engineering, 6068585, Japan*

[2]*Kyoto Institute of Technology, Division of Mechanodesign, 6068585, Japan*

ARTICLEINFO

Article history:
Received: 15 August, 2018
Accepted: 22 October, 2018
Online: 15 November, 2018

Keywords:
Attitude Control
Aerial Vehicle
Modeling

ABSTRACT

To gather information rapidly in disaster sites, a lot of search/rescue robots have been developed. It is difficult to correspond to the complicated environment composed of fields requiring various locomotion strategies, because most of these robots have only one type of locomotion device. To increase the available search routes under such conditions with the aim of gathering information more efficiently, we have previously proposed a legged aerial vehicle. The vehicle has tandem rotors to fly in the air and four legs to walk on the ground. The particular feature of this robot is that it has fewer actuators than the sum of those required for controlling a quadrupedal robot and a tandem-rotor helicopter individually. This paper presents modeling of the robot and development of an attitude control system that uses the leg motions. The behavior of the vehicle with the proposed attitude control is simulated using Multibody Dynamics (MBD) simulation software.

1 Introduction

Accidents and disasters such as building collapses, fires, earthquakes, and floods are responsible for massive loss of life. According to an interview survey about the Great Hanshin-Awaji Earthquake (also known as the Kobe Earthquake) of 1995, the most important contribution to rescue activities following such an event is rapid information gathering. Therefore, search robots are attracting attention as rapid means of gathering such information. In late years, a lot of researchers have been developed varied robots, and there has been remarkable progress in unmanned aerial and ground vehicles. For example, to save the electric power and to extend the inspection time on aged bridges, a UAV with a magnetic adsorption device was proposed and developed by Akahori et al [1]. The quadrotor helicopter-based UAV has an adsorption device using electro permanent magnets (EPM). And two cameras are set on the camera arms to carry out the close visual inspection.

Oliver et al. reported about a quadrotor helicopter with a tilt mechanism in [2]. The proposed quadrotor helicopter can fly horizontally while maintaining a steady attitude to conduct aerial inspection with stable condition. The development of unmanned multirotor helicopters has led to expand applications, such as (i) inspecting aerial power lines for maintenance, (ii) constructing platforms with which to rescue people, and (iii) undertaking construction on inaccessible areas [3]. On the ground, walking robots are able to cross irregular ground that is impossible to wheeled and crawler robots, leading to the development of experimental dynamic walking robot[4], [5]. Also, unmanned vehicles have been developed that can both terrestrial locomote and fly in the air. This is because in search scenarios it is often necessary to both fly quickly over extended areas and search carefully on the ground. Pratt and Leang developed the dynamic underactuated flying–walking (DUCK) robot, which combines a quadcopter helicopter with passive-dynamic legs to create a various system that can fly and walk [6].

To coordinate walking and flying abilities, in [7], the authors designed and constructed the legged air vehicle shown in Fig. 1. The frame, legs and the body are constructed from pipes and plates of carbon-fiber-reinforced polymer for high stiffness and low weight. Two brushless motor and propellers are equipped as the thruster. And four legs composed of servomotors and parallel link mechanisms are used for walking and

[*]Corresponding Author: Yoshiyuki Higashi, Matsugasaki, Sakyo-ku, Kyoto, +81-75-724-7364 & higashi@kit.ac.jp

https://dx.doi.org/10.25046/aj030627

controlling the attitude.

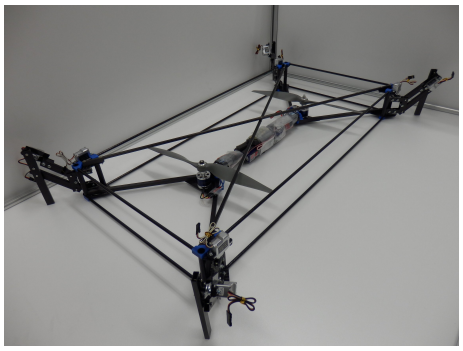


Figure 1: A prototype of the proposed legged aerial vehicle.

Table 1: Specifications of the legged aerial vehicle.

Length [mm]	702–870
Width [mm]	400–560
Height [mm]	120–200
Weight [g]	830

The flight attitude of the robot is controlled by changing of the center of gravity (CoG) caused by motion of the legs. The control using CoG can reduce the number of actuators than the sum of those required for controlling a quadrupedal robot and a tandem-rotor helicopter individually. With respect to flight control using change of CoG, there is a quadrotor helicopter controlled by an inverted pendulum. Miwa developed the quadrotor helicopter equipped an inverted pendulum on top of the body, and achieved control the attitude by using the tilt angle of the pendulum [8]. The present paper describes construction of mathematical models of the legs, rotors, and development of attitude control system based on them to simulate the motion of the robot in the air. And the behavior of the robot in the air is simulated using the constructed model and a physical model constructed in ADAMS. ADAMS is an analysis software to simulate the behavior of moving parts and distribution of loads and forces to the whole robot without solving the equations of motion analytically.

2 Model of a Leg and Parameter Identification

The inputs of the physical model constructed by ADAMS are forces such as the driving torques and thrusts of the legs and rotors, but the inputs of the actuator are voltage signals that depend on reference profiles of the leg angles and rotor speeds. Therefore, this section presents a mathematical model of the legs and identifies the relevant parameter values. The model is built to clarify the relation between the servomotor input and the drive torque. Figure 2 shows the servomotors for driving links around a roll axis and a yaw axis, and a parallel link mechanism of links.

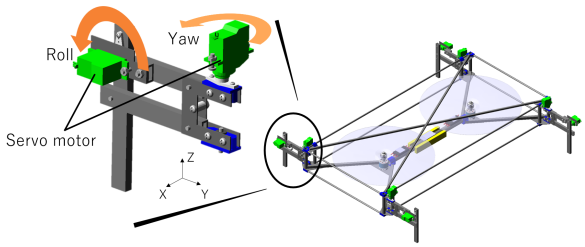


Figure 2: Servomotors and a parallel link mechanism of the leg

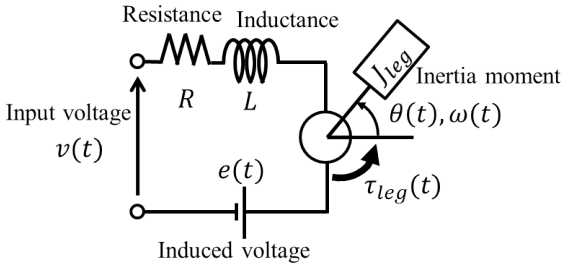


Figure 3: Electro-mechanical model of the direct current (DC) motor in one of the servomotors, including the effect of the moment of inertia, friction, and gravity.

We assumed that the leg could be modeled as a servomotor model that includes the influences of the moment of inertia, friction, and weight of the leg about each axis. Figure 3 shows the model of a direct current (DC) motor in servomotors. The driving torque of a the leg $\tau_{leg}(t)$ driving the leg is presented as the sum of the torques $\tau_m(t)$, $\tau_f(\omega(t))$, and $\tau_g(\theta(t))$. Here, $\tau_m(t)$, $\tau_f(\omega(t))$, and $\tau_g(\theta(t))$ are torque of the servomotor, friction, and gravity, respectively. Thus

$$\tau_{leg}(t) = \tau_m(t) + \tau_f(\omega(t)) + \tau_g(\theta(t)), \tag{1}$$

where t, $\omega(t)$ and $\theta(t)$ are time, the angular velocity and angle of the leg from a chosen reference.

The transfer function from the differential voltage to the motor $v_e(t)(= v(t) - e(t))$ to servomotor torque $\tau_m(t)$ the [9] is expressed as

$$\frac{T_m(s)}{V_e(s)} = \frac{K_t}{Ls + R}, \tag{2}$$

by using the coil inductance L, the motor resistance R and the torque coefficient K_t. And $T_m(s)$ and $V_e(s)$ are the Laplace transforms of $\tau_m(t)$ and $v_e(t)$.

The friction torque $\tau_f(\omega(t))$, which is a function of the angular velocity $\omega(t)$ of the leg, is given below[10]:

$$\tau_f(\omega(t)) = \begin{cases} -\tau_{act}(t) & \omega(t) = 0 \ \wedge \ |\tau_{act}(t)| < \tau_s \\ -\tau_s \mathrm{sgn}(\tau_{act}(t)) & \omega(t) = 0 \ \wedge \ |\tau_{act}(t)| \geq \tau_s \\ \tau_a(\omega(t)) & \text{otherwise} \end{cases} \tag{3}$$

$$\tau_a(\omega(t)) = -\mathrm{sgn}(\omega(t))\{a_1 \exp(-\mathrm{sgn}(\omega(t))\omega(t)) + \mathrm{sgn}(\omega(t))a_2 + a_3\omega(t)\}, \tag{4}$$

where τ_s is the static friction torque, $\tau_{act}(t)$ is the sum of $\tau_m(t)$ and the gravity torque $\tau_g(t)$, and a_1, a_2, and a_3 ($a_1 + a_2 = \tau_s$) are positive. Therefore, the Laplace transforms of the leg angle $\Theta(s)$ is presented as

$$\Theta(s) = \frac{T_{leg}(s)}{J_{leg}s^2}, \tag{5}$$

Hrere, T_{leg} and J_{leg} are the Laplace transforms of $\tau_{leg}(t)$ and the moment of inertia of links and servomotors.

The servomotor controller is taken to be a proportional–derivative (PD) controller [9] as given by

$$v(t) = K_{sp}(\theta_{ref}(t) - \theta(t)) + K_{sd}(\omega_{ref}(t) - \omega(t)), \tag{6}$$

where K_{sp} and K_{sd} are the proportional gain and differential gain, respectively. $\theta_{ref}(t)$ is the reference input of the leg angle.

In generally servomotors are controlled by PD controller to earn rapid response. The controller does not include an integral (I) controller. Therefore the measurement angle has slight angle error to the reference as shown in Figs. 4 and 5. The drive angle of a DC motor in the servomotor is measured by an installed potentiometer or a rotary encoder. Measured value has some electric noise, however, the noise does not cause terrible error because these signal passes a low pass filter. The unknown parameters A, B and C are estimated using the Levenberg–Marquardt method which is one of the iterative technique that locates the minimum of a function that is expressed as the sum of squares of nonlinear functions. To identify the parameters, the response of the leg about each axis was measured by using following the three steps below:

1. Attach two light-emitting diodes (LEDs) to the leg (see Figs. 4 and 5).
2. Record the motion of the leg to a video (30 fps) and the reference angle to a micro computer.
3. Apply image processing to each frame of the video to calculate the coordinates of each LED.
4. From those coordinates, calculate the angle of the leg for each frame.

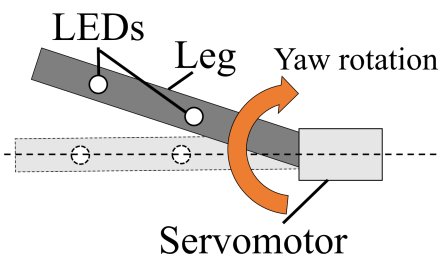


Figure 4: Leg motion in angle measurement experiment for the yaw rotation using LEDs.

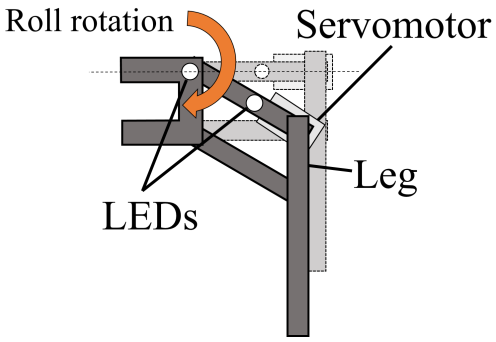


Figure 5: Leg motion in angle measurement experiment for the roll rotation using LEDs.

We determine the parameter values by minimizing the squared difference between the measurements and the simulation values using the models shown in Figs 4 and 5. Here, the simulation frequency was 1 kHz. Table 2 lists the identified parameter values, and Figs. 6 and 7 show the simulation results for each axis accompanied by the respective measurements. These simulation results shows that the behavior of the leg model is close to the measured behavior.

Table 2: Result of Parameter Identification

Symbol	Identified value
K_t	0.2016
a_3	0.035
K_{sp}	0.912
K_{sd}	0.005
a_2	0.070

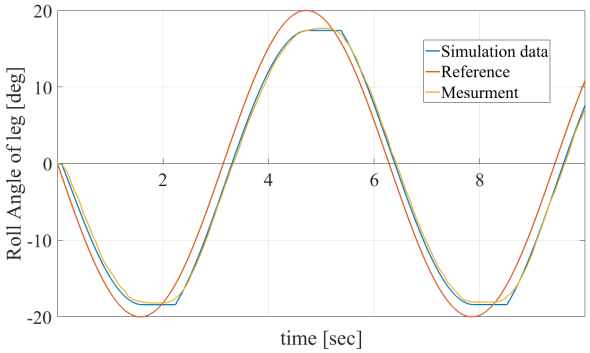


Figure 6: Simulation results of roll angle of the leg. Blue line is simulation result using identified parameters. Mearment data is measured roll angle through the experiment shown in Fig. 5.

3 Model of Rotor

We explain the model of a rotor composed of a propeller and a brushless motor assumed as a DC motor in this section. The brushless motor is assumed as a DC motor. Thus, as with Eq. (2), the transfer function from the differential voltage to the motor torque is given as

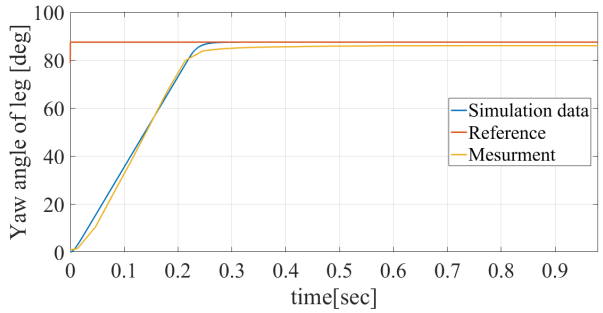


Figure 7: Simulation results of roll angle of the leg. Blue line is simulation result using identified parameters. Mearment data is measured yaw angle through the experiment shown in Fig. 4.

$$\frac{T_{bm}(s)}{V_{be}(s)} = \frac{K_{bt}}{L_b s + R_b}, \tag{7}$$

where each variable corresponds to the respective one in Eq. (2).

The thrust $T(t)$ generated by the propeller and the counter-torque $\tau_c(t)$ acting on the propeller due to the air are known to be proportional to the square of the rotor velocity in the steady state. Therefore, $T(t)$ and $\tau_c(t)$ are calculated using approximation curves based on the data sheet of the propeller on the manufacturer APC's web site [11]. Those thrust and torque are given as

$$T(t) = 3.026 n^2(t), \tag{8}$$
$$\tau_c(t) = 5.039 n^2(t), \tag{9}$$

where $n(t)$ is the rotation velocity in units of revolutions per minute (rpm); Eqs. (8) and (9) are plotted in Figs. 8 and 9, respectively. Moreover, the relationship between the rotation velocity $\psi(t)$ in units of radians per second (rad/s) and the torque $\tau_p(t)(= \tau_{bm}(t) - \tau_c(t))$ is

$$\frac{\dot{\Psi}(s)}{T_p(s)} = \frac{1}{J_p s}, \tag{10}$$

where J_p is the inertia torque of the propeller, and $\dot{\Psi}(s)$ and $T_p(s)$ are the Laplace transforms of $\dot{\psi}(t)$ and $\tau_p(t)$. The response of the rotor is simulated based on these equations.

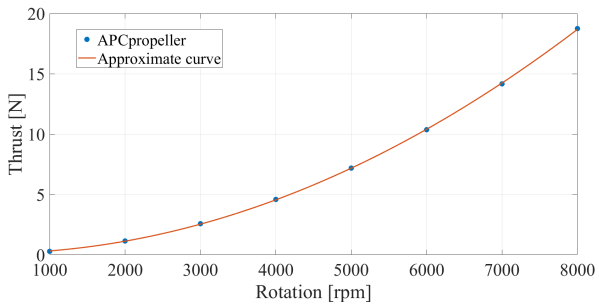


Figure 8: Thrust data from manufacturer's site [11] and approximation curve.

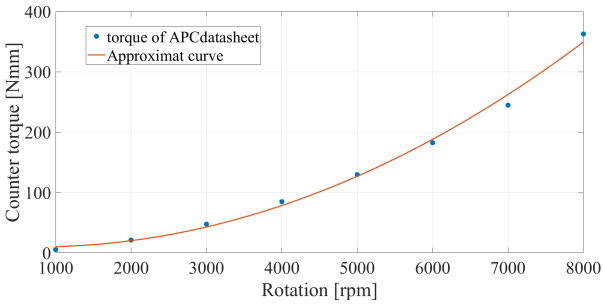


Figure 9: Counter-torque data from manufacturer's site [11] and approximation curve.

4 Attitude Control using the Leg Motion

4.1 Modeling of CoG and Leg Angles

To develop a control system for CoG control, we clarify relation between the CoG and the leg angles. The variables and link names are defined as shown in Figs. 10 and 11. The masses of links B, U, L, and F are m_b, m_u, m_l, and m_f, respectively, and those of the leg and the vehicle are m and M, respectively. The matrix $\boldsymbol{\theta} \in \mathbb{R}^{2\times 4}$ of legs angles is defined as

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_{FLy} & \theta_{FRy} & \theta_{RLy} & \theta_{RRy} \\ \theta_{FLr} & \theta_{FRr} & \theta_{RLr} & \theta_{RRr} \end{bmatrix} \tag{11}$$
$$= \begin{bmatrix} \boldsymbol{\theta}_y \\ \boldsymbol{\theta}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\theta}_{FL} & \boldsymbol{\theta}_{FR} & \boldsymbol{\theta}_{RL} & \boldsymbol{\theta}_{RR} \end{bmatrix}. \tag{12}$$

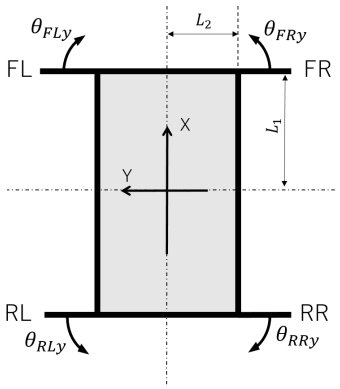


Figure 10: Definition of leg angles around the yaw axis.

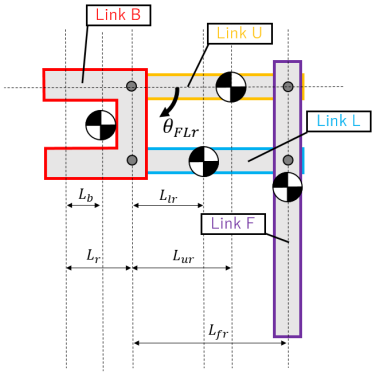


Figure 11: Definition of angles and link parameters on the front-left (FL) leg.

By the definitions, the robot CoG $\boldsymbol{G}(\boldsymbol{\theta}) = \{G_X, G_Y\}^T$ is expressed following Eq. (13) in terms of the various leg CoGs, namely $\boldsymbol{G}_{FL}(\boldsymbol{\theta}_{FL})$, $\boldsymbol{G}_{FR}(\boldsymbol{\theta}_{FR})$, $\boldsymbol{G}_{RL}(\boldsymbol{\theta}_{RL})$, and $\boldsymbol{G}_{RR}(\boldsymbol{\theta}_{RR})$:

$$\begin{aligned}\boldsymbol{G}(\boldsymbol{\theta}) = \frac{m}{M}(&\boldsymbol{G}_{FL}(\boldsymbol{\theta}_{FL}) + \boldsymbol{G}_{FR}(\boldsymbol{\theta}_{FR}) \\ &+ \boldsymbol{G}_{RL}(\boldsymbol{\theta}_{RL}) + \boldsymbol{G}_{RR}(\boldsymbol{\theta}_{RR})).\end{aligned} \tag{13}$$

Each leg CoG is determined from the respective leg angle. The CoG of the front-left leg (FL) is given by

$$\boldsymbol{G}_{FL}(\boldsymbol{\theta}_{FL}) = \begin{bmatrix} \frac{1}{m}(mL_1 + ((m_u + m_l + m_f)L_r + m_b L_b \\ +(m_u L_{ur} + m_l L_{lr} + m_f L_{fr})\cos\theta_{FLr})\sin\theta_{FLy}) \\ \\ \frac{1}{m}(mL_2 + ((m_u + m_l + m_f)L_r + m_b L_b \\ +(m_u L_{ur} + m_l L_{lr} + m_f L_{fr})\cos\theta_{FLr})\cos\theta_{FLy}) \end{bmatrix}. \tag{14}$$

The vehicle CoG is calculated from Eqs. (13) and (14) and the various leg angles.

However, the reference leg angles cannot be determined uniquely from the reference CoG because of $\boldsymbol{\theta} \in \mathbb{R}^{2\times4}$ and $\boldsymbol{G} \in \mathbb{R}^2$. Assuming that leg motion around the yaw axis and roll axis cause movement of the CoG along X and Y axis each, the conditions for the leg yaw angles are

$$\begin{cases} \theta_{RLy} = \theta_{RRy} = 0, \\ \boldsymbol{G}_{FLX}(\theta_{FLy}, \hat{\theta}_{FLr}) = \\ \quad \boldsymbol{G}_{FRX}(\theta_{FRy}, \hat{\theta}_{FRr}) \quad (G_X \geq 0) \\ \\ \theta_{FLy} = \theta_{FRy} = 0, \\ \boldsymbol{G}_{RLX}(\theta_{RLy}, \hat{\theta}_{RLr}) = \\ \quad \boldsymbol{G}_{RRX}(\theta_{RRy}, \hat{\theta}_{RRr}) \quad (G_X < 0) \end{cases} \tag{15}$$

by using the estimated leg angle matrix $\hat{\theta}$. And those for the leg roll angles are

$$\begin{cases} \theta_{FLr} = \theta_{RLr} = 0, & (G_Y \geq G_Y(\hat{\boldsymbol{\theta}}_y, 0)) \\ \theta_{FRr} = \theta_{RRr} = 0, & (G_Y < G_Y(\hat{\boldsymbol{\theta}}_y, 0)). \end{cases} \tag{16}$$

Also, when $G_Y \geq G_Y(\hat{\boldsymbol{\theta}}_y, 0)$ is satisfied, the additional conditions are

$$\begin{cases} G_{FRY}(\hat{\theta}_{FRy}, \theta_{FRr}) = G_{RRY}(\hat{\theta}_{RRy}, \theta_{RRr}) \\ \quad (\{G_{FRY}(\hat{\theta}_{FRy}, 0) \geq G_{RRY}(\hat{\theta}_{RRy}, 0)\} \\ \quad \wedge\{G_Y \geq \frac{m}{M}(G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad + G_{RLY}(\hat{\theta}_{RLy}, 0) + 2G_{RRY}(\hat{\theta}_{RRy}, 0))\}) \\ \\ \theta_{RRr} = 0, \\ G_{FRY}(\hat{\theta}_{FRy}, \theta_{FRr}) = \frac{M}{m}G_Y - G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad - G_{RLY}(\hat{\theta}_{RLy}, 0) - G_{RRY}(\hat{\theta}_{RRy}, 0) \\ \quad (\{G_{FRY}(\hat{\theta}_{FRy}, 0) \geq G_{RRY}(\hat{\theta}_{RRy}, 0)\} \\ \quad \wedge\{G_Y < \frac{m}{M}(G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad + G_{RLY}(\hat{\theta}_{RLy}, 0) + 2G_{RRY}(\hat{\theta}_{RRy}, 0))\}) \\ \\ G_{FRY}(\hat{\theta}_{FRy}, \theta_{FRr}) = G_{RRY}(\hat{\theta}_{RRy}, \theta_{RRr}) \\ \quad (\{G_{FRY}(\hat{\theta}_{FRy}, 0) < G_{RRY}(\hat{\theta}_{RRy}, 0)\} \\ \quad \wedge\{G_Y \geq \frac{m}{M}(G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad + G_{RLY}(\hat{\theta}_{RLy}, 0) + 2G_{FRY}(\hat{\theta}_{FRy}, 0))\}) \\ \\ \theta_{FRr} = 0, \\ G_{RRY}(\hat{\theta}_{RRy}, \theta_{RRr}) = \frac{M}{m}\bar{G}_Y - G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad - G_{RLY}(\hat{\theta}_{RLy}, 0) - G_{FRY}(\hat{\theta}_{FRy}, 0) \\ \quad (\{G_{FRY}(\hat{\theta}_{FRy}, 0) < G_{RRY}(\hat{\theta}_{RRy}, 0)\} \\ \quad \wedge\{G_Y < \frac{m}{M}(G_{FLY}(\hat{\theta}_{FLy}, 0) \\ \quad\quad + G_{RLY}(\hat{\theta}_{RLy}, 0) + 2G_{RRY}(\hat{\theta}_{RRy}, 0))\}). \end{cases} \tag{17}$$

When another condition is to be satisfied, the additional conditions are given with suffixes representing the right and left sides of the robot. Thus, the reference leg angles are determined based on the relation between a leg CoG and this angle in the conditions.

4.2 Calculation of Reference Position of CoG

The reference CoG for attitude control is calculated from the error in the attitude angle trajectory. In the first step, the pitch-roll-yaw torque around the CoG acting the entire aircraft $\boldsymbol{\tau}_{CoG}(t) = \{\tau_p(t), \tau_r(t), \tau_y(t)\}$ is calculated from a PD controller shown below:

$$\boldsymbol{\tau}_{CoG}(t) = \boldsymbol{K}_p \circ (\boldsymbol{\phi}_{ref}(t) - \boldsymbol{\phi}(t)) + \boldsymbol{K}_d \circ (\dot{\boldsymbol{\phi}}_{ref}(t) - \dot{\boldsymbol{\phi}}(t)) \tag{18}$$

by using the pitch-roll-yaw $(\phi_p(t), \phi_r(t), \phi_y(t))$, $\boldsymbol{\phi}(t) \in \mathbb{R}^3$, and their reference $\boldsymbol{\phi}(t)_{ref} \in \mathbb{R}^3$. Where $\boldsymbol{K}_p$, $\boldsymbol{K}_d \in \mathbb{R}^3$ are the proportional gain and differential gain, $\dot{*}$ is differential of the $*$, and $\circ$ is Hadamard product. In the second step, the thrusts and the center of thrust (CoT), which is the position of the torque around the CoG balancing caused by the thrust, is estimated from the input values using the rotor model. The estimated thrusts and CoT are represented by $\hat{\boldsymbol{T}}(t) = \{\hat{T}_1(t), \hat{T}_2(t)\}^T$ and $\hat{\boldsymbol{CoT}}(t) = \{\hat{CoT}_X(t), \hat{CoT}_Y(t)\}^T$, respectively. The final step determines the reference CoG $\boldsymbol{G}_{ref}(t) = \{X_{ref}, Y_{ref}\}^T$ from the torque, the thrust, and the CoT given by the first and second steps by the

following equation:

$$\boldsymbol{G}_{ref}(t) = \frac{1}{\hat{T}_1(t) + \hat{T}_2(t)} \{\tau_p(t), \tau_r(t)\}^T - \boldsymbol{C\hat{o}T}(t). \quad (19)$$

Meanwhile, reference rotor speed is determined by $\tau_y(t)$ and the reference of the sum of the thrusts is determined using Eqs. (10) and (9).

5 Attitude Control Simulation

The motion of the robot was simulated based on the physical model constructed by ADAMS and the control system constructed by a numerical analysis software MATLAB. Simulation parameters are the same values to those of the prototype. The sampling frequency of simulations was 1 kHz. And Table 4 shows reference angles in the simulations.

Figures 12 and 13 show the pitch angle for simulation pattern 1 and the roll angle for simulation pattern 2. Fig. 14 shows the result of the yaw angle for simulation pattern 3. Results in Figs. 12, 13, and 14 corresponded with the respective reference lines, and showed that the developed control system was effective through the simulations.

Figures 15 - 19 show detail of the pitch, the roll and the yaw angle in several reference angles including large reference. When the pitch reference is 20 deg shown in Fig. 15, the pitch angle follows the reference rapidly. although the roll angle changs like a sine wave in this simulation,the control paformance is enough effective because the range is very narrow between -0.5 and 0.5deg. In case of the roll reference is 20 deg, the pitch angle did almost not change as shown in Fig. 16. The reason that the change of the pitch angle in Fig. 16 is smaller than the change of the roll in Fig. 15, is difference of the moment of inertia and variable range of the CoG

The moment of inertia of the robot around the pitch axis is larger than one around the roll axis. And the range of change of CoG to control the pitch angle is also wider than one of the roll control. Therefore the stability of pitch angle is better than the performance of the roll angle.

The effect appeared in the yaw angle too. However the change of except for target angle is very small, and these results showed validness of the developed controller. Especially the roll control worked in large reference 25 deg (Fig. 17) which was enough for flight control. The yaw control also generated slight change in other axes when reference are low (Fig. 18) and high (Fig. 19).

Table 3: Proportional–derivative (PD) controller gains

	K_p	K_d
Pitch	1.0	0.5
Roll	2.0	1.3
Yaw	0.1	0.55

Table 4: Reference of attitude angles in control simulations

	Pitch [deg]	Roll [deg]	Yaw [deg]
Pattern 1	5.0	0.0	0.0
Pattern 2	0.0	5.0	0.0
Pattern 3	0.0	0.0	20.0
Pattern 4	20.0	0.0	0.0
Pattern 5	0.0	20.0	0.0
Pattern 6	0.0	25.0	0.0
Pattern 7	0.0	0.0	10.0
Pattern 8	0.0	0.0	30.0

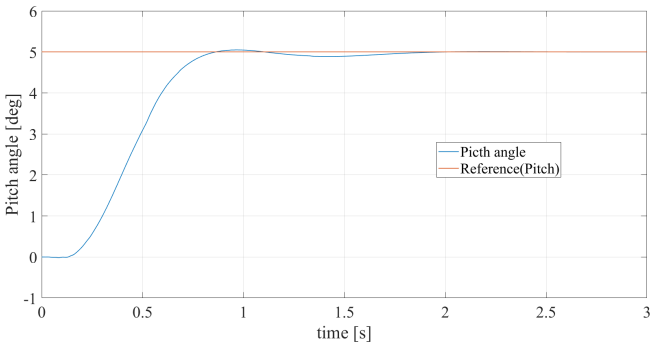


Figure 12: Step response of pitch-angle controller (pattern 1, reference pitch:5.0, roll:0.0, yaw:0.0).

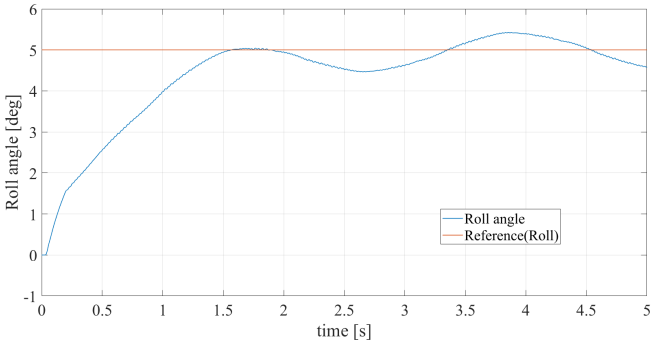


Figure 13: Step response of roll-angle controller (pattern 2, reference pitch:0.0, roll:5.0, yaw:0.0).

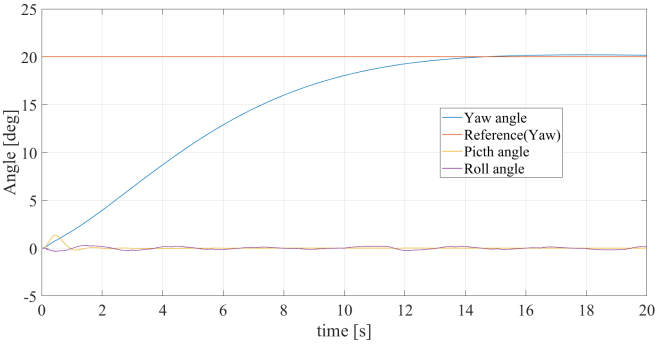


Figure 14: Step response of yaw-angle controller (pattern 3, reference pitch:0.0, roll:0.0, yaw:20.0)

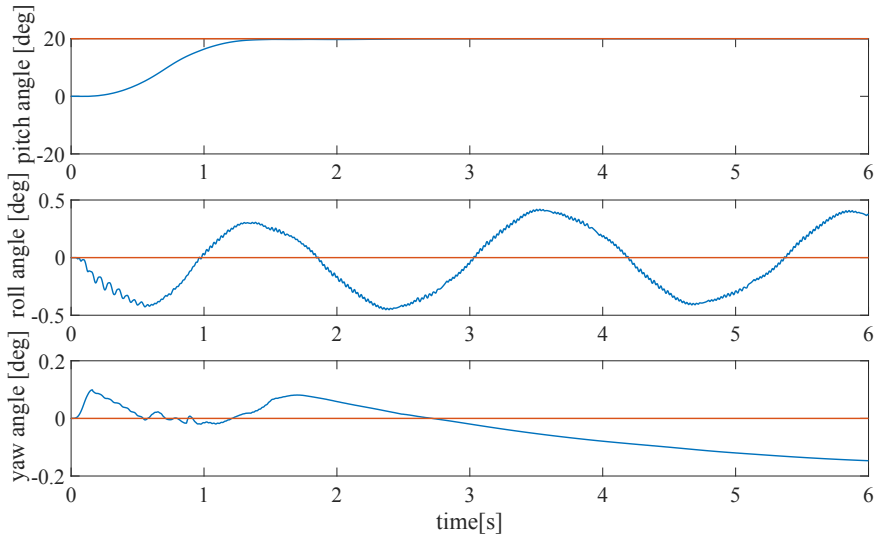


Figure 15: Response of three axes (pattern 4, reference pitch:20.0, roll:0.0, yaw:0.0)

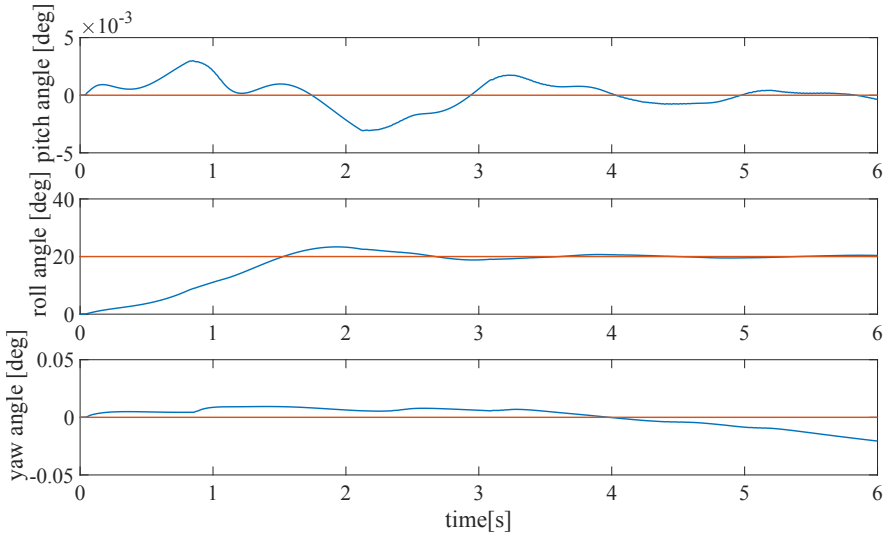


Figure 16: Response of three axes (pattern 5, reference pitch:0.0, roll:20.0, yaw:0.0)

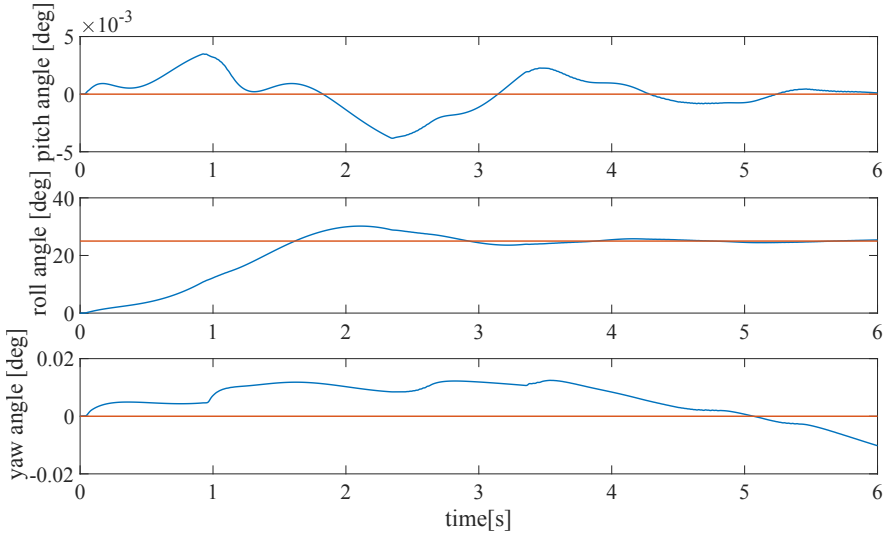


Figure 17: Response of three axes (pattern 6, reference pitch:0.0, roll:25.0, yaw:0.0)

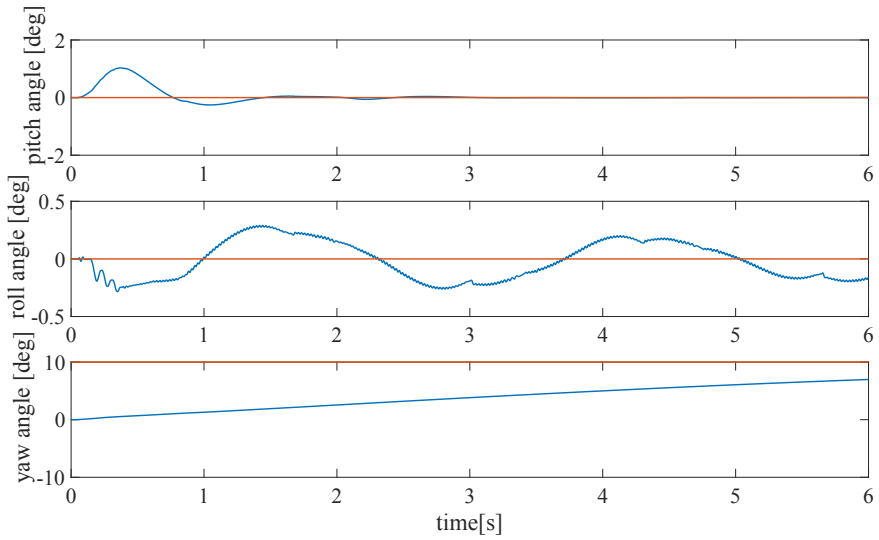


Figure 18: Response of three axes (pattern 7, reference pitch:0.0, roll:0.0, yaw:10.0)

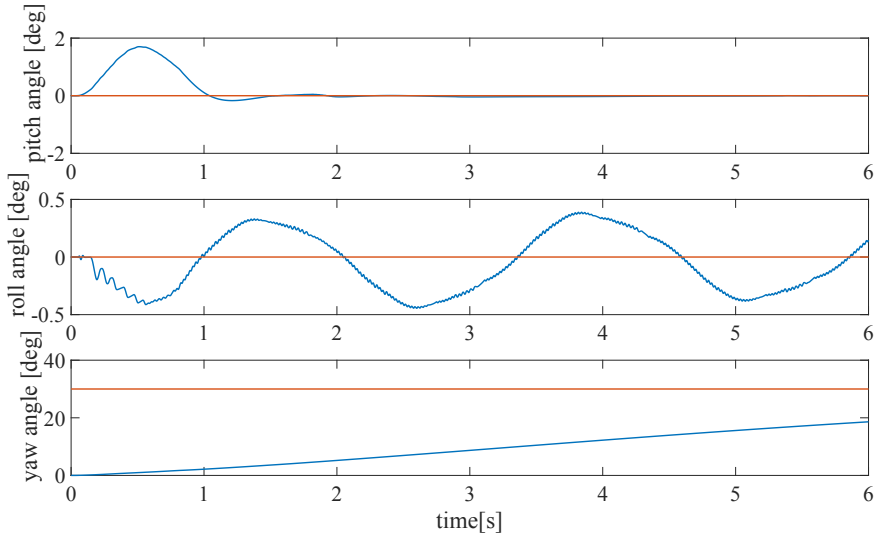


Figure 19: Response of three axes (pattern 8, reference pitch:0.0, roll:0.0, yaw:30.0)

6 Conclusions

The paper has described modeling of the leg and the rotor including the propeller, and a control method using the change of the CoG caused by the leg motion. Flight attitude was simulated for some reference angles by the developed control system. These simulation results shows that (i) response of the leg model corresponded to the experimental data and (ii) the attitude angles of the robot could follow their references using the attitude control system. Moreover the difference of moment of inertia and range of CoG in each axis appeared in the results, it was shown that the constructed model could reflect the characteristic of the robot. However, the results also showed that the transient response is not yet fast enough. Particularly convergence of the yaw angle for the yaw reference was slow, even though the yaw angle could follow the reference at last. Future work will involve implementing the control system on the developed legged aerial robot to evaluate the control performance.

References

[1] S. Akahori, Y. Higashi, A. Masuda, Development of an Aerial Inspection Robot with EPM and Camera Arm for Steel Structures, Proceedings of the International Conference, pp. 3546-3549, 2016.

[2] Russell Oliver, Sui Yang Khoo, Michael Norton, Scott Adams, Abbas Kouzani, Development of a Single Axis Tilting Quadcopter, Proceedings of the Proceedings of the International Conference, pp. 1851-1854, 2016.

[3] A.E Jimenez-Cano, J. Braga, G. Heredia, A. Ollero, Aerial Manipulator for Structure Inspection by Contact from the Underside, Proceedings of the International Conference on Intelligent Robots and Systems, pp. 1879-1884, 2015.

[4] S. Kitano, S. Hirose, G. Endo, Design and Development of Quadruped Robot TITAN-XIII, Journal of Japan Society for Design Engineering, volume 51, number 12,pp. 875-884, 2016.

[5] M. Raibert, K. Blankespoor, N. Gabriel, R. Playter, BigDog, the Rough - Terrain Quadruped Robot, Proceedings of the 17th IFAC World Congress, Vol, 41, Issue 2, pp. 10822-10825, 2008.

[6] C. J. Pratt and K. K. Leang, Dynamic Underactuated Flying-Walking (DUCK) Robot, Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 3267-3274, 2016.

[7] Y. Higashi and R. Okada, Suggestion of Legged Air Vehicle for Locomotion on ground and in air, Proceedings of the 8th International Conference on Intelligent Unmanned Systems, pp. 228-231, 2012.

[8] M. Miwa, S. Kunou, S. Uemura, A. Imamura, H. Niimi, Attitude Control of Quad-Rotor Helicopter with COG Shift, Journal of JSEM, vol. 13, Special Issue, pp. s102-s107, 2013.

[9] Ishikawa M., Kitayoshi R., Wada T., Maruta I., Sughie S., Modeling of R/C Servo Motor and Application to Underactuated Mechanical Systems, Transactions of the Society of Instrument and Control Engineers Vol.46, No.4, pp. 237-244, 2010.

[10] B.Armstrong, Friction:Experimental Determination,Modeling and Compensation, Proceedings of the 1988 IEEE International Conference on Robotics and Automation, p.1422, 1988.

[11] J. J. More, The Levenberg–Marquardt algorithm: Implementation and theory Proceedings of Conference on numerical analysis, CONF-770636-1, 1997.

[12] APC Propeller Performance Data, http://www.apcprop.com/v/downloads/PERFILES_WEB/datalist.asp, 2017/01/27

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 207-215 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering 2018

ASTESJ

ISSN: 2415-6698

An Experimental Investigation of Reconfigurable UWB Modified Octagonal Microstrip Monopole Patch Antenna with Switchable and Tunable Band-Notched Characteristic

Ruchi Paliwal[*,1], Ruchi Paliwal[1], Rajesh Kumar Singh[2], Shiban Kishen Koul[3]

[1]*Dept. of Electronics and Comm. Engg., JSS Academy of Technical Education, Noida (UP), 201301, India*

[2]*Bharti School of Telecomm. Tech. and Management, IIT Delhi, 110016, India*

[3]*CARE, IIT Delhi, 110016, India*

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 24 September, 2018
Online: 05 October, 2018

Keywords:
Reconfigurable antennas
PIN diodes
Ultra-wideband

ABSTRACT

An experimental investigation of reconfigurable ultra-wideband (UWB) modified octagonal microstrip monopole patch antenna with switchable and tunable band-notched characteristic is presented in this paper. Notched-bands are obtained by connecting or disconnecting the open shunt stubs. The switching is achieved by using PIN diodes. Additionally, more notch-bands are created by adding additional stubs to the original design and tuning is achieved by employing a varactor diode in the circuit. The proposed antenna has more freedom to suppress the interfering bands because of tuning nature. The proposed designs are simulated, fabricated and then tested. The proposed structures are simple in geometry and useful to operate in the rich electromagnetic scattering environment.

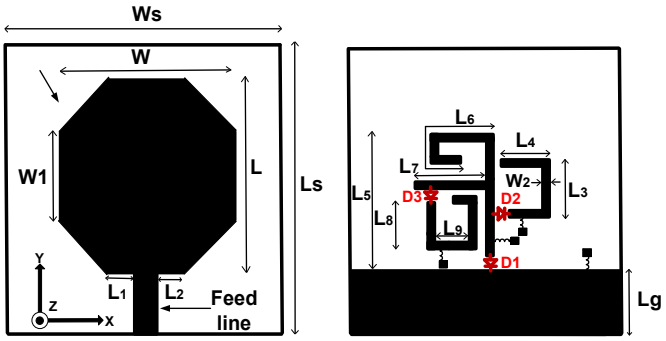
1. Introduction

In today's world, wireless communication system becomes an essential need for communication. It requires an antenna that must be effective to change its characteristics in reversible manner and it can be utilized for many applications and functions. With the help of transformable antenna, we can replace multiple antennas by a single antenna element which can utilize the complete unlicensed ultra-wideband (UWB) [1]. On this ground, Planar UWB antennas are in great demand [2-7]. While using the entire bandwidth of UWB, other signals that exist in wireless communication system like Worldwide Interoperability for Microwave Access (WiMAX), wireless local area network (WLAN), X-band downlink, X-band satellite communication link, X-band aeronautical radio navigation, etc., operating around 3.5, 5, 5.8, 7.5, and 9 GHz band, respectively, can cause interference. To mitigate this issue, band-notched UWB antennas have been designed that can filter out interfering frequency and improved selectivity. The band-notched UWB antennas have been presented in the past few years.

By inserting a proper slots or slits with different shapes in a radiating element, notched-band can be achieved [3]. In [4, 5], one notch is created while in [7, 8, 9, 10, 11, 12] two notched-bands are obtained. In [13], notched-band has been created by using electromagnetic band-gap (EBG) structure. In [14], one notch band is achieved from CPW-fed monopole antenna, but the structure is compact. Wide notch bands have also been achieved in [15]. Besides this, antenna with static feature will become obsolete for next generation wireless systems. Antenna with reconfigurable feature can be a good candidate for these systems. Reconfigurable antennas have been created by connecting and disconnecting different conducting elements to the main geometry with the help of electrical switches like radio frequency micro-electro-mechanical systems (RF MEMS), PIN diodes or by other methods. Reconfigurability can also be achieved with optical switch or by using material change [16]. A reconfigurable UWB antenna with tuning notched-bands could be advantageous in the future wireless communication systems. Variable notched frequency bands have been achieved in [17], [18]. Some compact UWB monopole antennas reported in the literature can be integrated in a compact system and save space [19], [20]. A reconfigurable UWB modified octagonal monopole microstrip patch antenna with switchable and tunable notched-bands is discussed in this paper. The work proposed in this paper is an extended work originally presented in [21]. The proposed antenna has more freedom to suppress the selected bands.

[*]Ruchi Paliwal, JSS Academy of Technical Education, Noida (UP), 201301, India, 0120-2400104, ruchipaliwal@jssaten.ac.in

The paper starts with the introduction. Section 2 describes the design of a reconfigurable UWB monopole patch antenna. Section 3 discusses the modified design to create more notched-bands by adding thin stubs. Section 4 presents the tuning in the structure by using a varactor diode and finally concludes the paper with results and discussions.



(a) (b)

Figure 1. Geometrical model of the modified octagonal patch antenna (a) front side (b) back side.

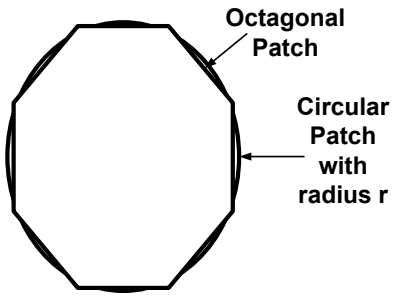


Figure 2. Geometry of a conventional octagonal patch and circular patch

Table 1: Dimensions of the proposed reconfigurable UWB monopole antenna

Parameter	Value (mm)	Parameter	Value (mm)
L	25.50	L8	04.65
L1	04.00	L9	03.30
L2	04.00	Ls	42.00
L3	03.40	W	22.50
L4	04.23	W1	14.50
L5	09.30	W2	00.60
L6	05.50	Ws	40.00
L7	06.60		

2. Reconfigurable UWB monopole patch antenna design and its operation

2.1. UWB antenna design

The geometrical model of the antenna is illustrated in Figure 1. The antenna is designed on FR4 substrate with relative permittivity of 4.4. The height of the substrate is 1.6 mm. The proposed antenna consists of modified octagonal monopole microstrip patch etched on the upper side and partial ground plane on the lower side of the substrate. Thin metal stubs are etched on the back side of the patch to create notches in the reflection coefficient plot and this is achieved by integrating PIN diodes in the circuit as shown in Figure 1(b). The conventional octagonal (with equal side length) patch is designed by taking its size equivalent to a circular patch as shown in Figure 2. The dominant mode is TM_{110} (TM_{mnp}), the resonant frequency can be calculated as

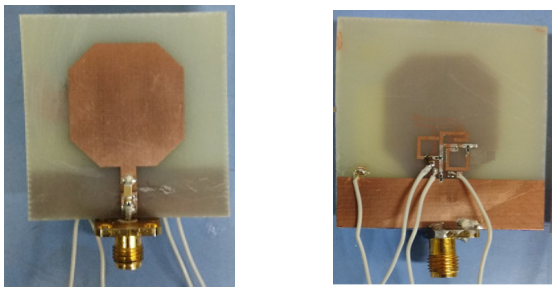
$$f_r = \frac{1.8412\ c}{2\pi r\sqrt{\varepsilon_r}} \quad (1)$$

Where

c = Speed of light,

r = Radius of the circular patch,

εr = Relative permittivity.



(a) (b)

Figure 3. Photograph of a modified octagonal monopole microstrip patch antenna (a) front side (b) back side.

The radius is calculated as 14.05 mm by equating the area of both structures (circular and octagonal patch). The resonant frequency of the dominant mode is 2.94 GHz. The proposed octagonal patch is modified by varying the edges and using partial ground plane and achieved impedance matching throughout the band (3.1-10.6 GHz). The proposed antenna is excited with a 50 Ω line. There are total of three PIN diodes named as D1, D2, and D3 are connected at different places to achieve different notches. PIN diodes are turned ON with the help of biasing circuits at the back side of antenna. The dimensions of the circuit are listed in Table 1. Photograph of the fabricated circuit is shown in Figure 3.

2.2. Operation and results

Three PIN diodes are used to get various states from the circuit. Simulations are done in an EM simulator (Computer Simulation Technology software) [22]. The equivalent model of a PIN diode used in the simulations is shown in Figure 4. The values of Rp, Cp and Rs are 40 kΩ, 0.03 pF and 5 Ω, respectively. The parasitic series inductance is ignored in the simulations. The equivalent model is extracted from the datasheet [23]. Inductors of 56 nH are used to bias the PIN diodes. A capacitor of 100 pF is used in the circuit to block dc. Input impedance of the proposed antenna is plotted in Figure 5. From

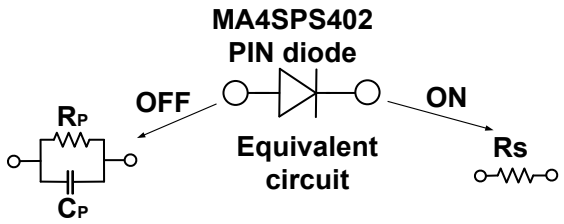
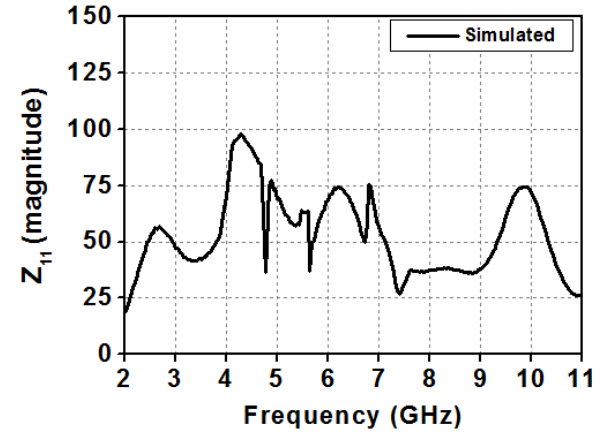
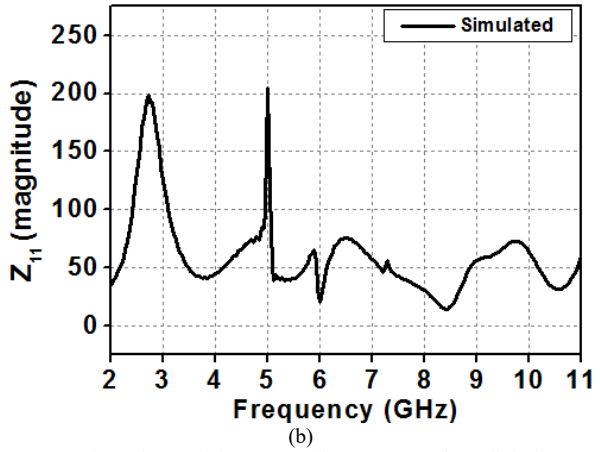


Figure 4. The equivalent circuit model of the PIN diode.



(a)



(b)

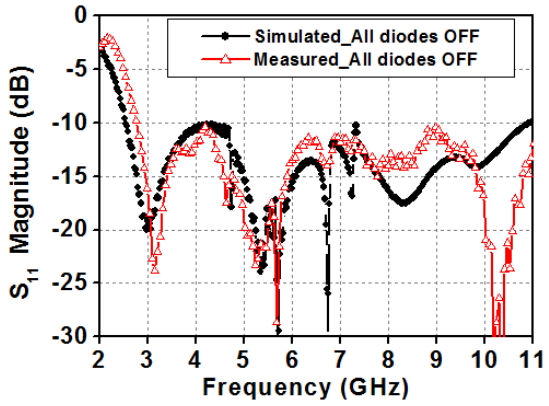
Figure 5. Input impedance of the proposed antenna (a) when all diodes are OFF, (b) diode D1is ON.

Figure 5(a), it is clear that the input impedance lies in the range from 26 Ω to 96 Ω which corresponds to S_{11} < -10 dB. It is matched in the whole UWB band. Figure 5(b) shows the input impedance of 203 Ω (S_{11} = - 4.37 dB) at 5.05 GHz and 13 Ω (S_{11} = - 4.62 dB) at 8.25 GHz, which creates two notches. When all diodes are OFF, the impedance bandwidth is achieved in the range from 2.8 to 11.5 GHz as shown in Figure 6(a). In the second state, diode D1 is turned ON (L5 and L6 stubs are connected to the ground plane); two notched-bands are created around at 5.05 and 8.25 GHz as shown in Figure 6(b). In the third state, all three diodes are ON (all stubs are connected), the notched-bands are shifted to some other bands with centre frequencies of 5.5 and 8 GHz as shown in Figure 6(c) but in this case, the impedance bandwidth is reduced. At 3.1 GHz, the magnitude of reflection coefficient is below -10 dB which is not acceptable.

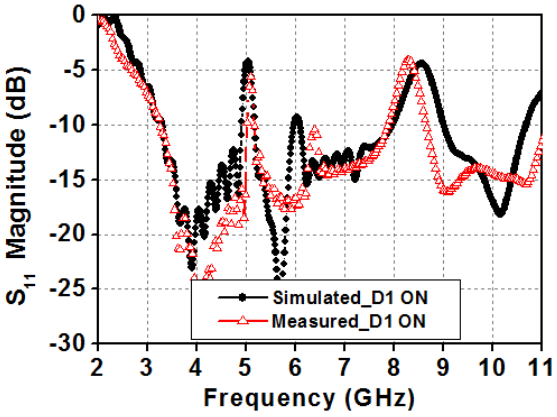
Further the design is modified to get good impedance match across the whole UWB band except the notched-bands.

3. Modified reconfigurable UWB antenna

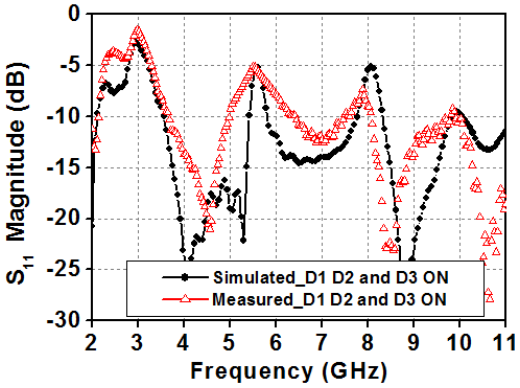
The modified structure is shown in Figure 7. The geometry of the top side is same as the previous one; the only change is done on the back side (partial ground plane side). Photograph of the printed design is shown in Figure 8. Two additional stubs are connected in this modified geometry to get more notched-bands. There are 4 PIN diodes in the circuit. Here, measurements are done by taking PIN diode as a metal strip of 0.2 mm × 0.2 mm. Input impedance is plotted in Figure 9. It is perfectly matched in the range 2.8–11.4 GHz. When all four diodes are OFF, the reflection coefficient is below -10 dB across the whole band (3.1-10.6 GHz) as depicted in Figure 10(a). UWB characteristic is achieved from the circuit without turning the diodes ON. In state 2, diode D1 is ON and others are OFF, two notched-bands are created as show in Figure 10(b). In state 3, diodes D1 and D2 are ON, one notched-band is present (around 8.5 GHz) and other notched-band (around 5 GHz) is suppressed because when we connect D2, one more shunt stub (L3 + L4) is connected to the circuit, this open ended stub (length = λ/4) suppress the 5 GHz band as shown in Figure 10(c). In state 4, diodes D1 and D3 are ON, notches are shifted as shown in Figure 10(d), because of change in the effective reactance value. In state 5, diodes D1 and D4 are ON; again, it suppresses the 5 GHz band and obtains the wide notched-band at 7 GHz (center frequency) as shown in Figure 10(e). In the last case (state 6), all diodes are ON, two notched-bands at 3.5 and 7.25 GHz are created as shown in Figure 10(f).



(a)



(b)



(c)

Figure 6. Reflection coefficients of the proposed antenna.

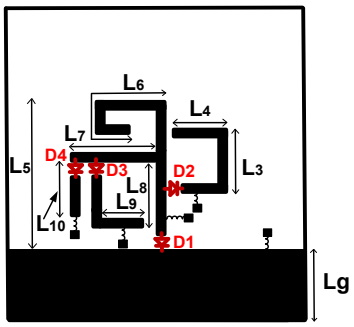


Figure 7. Proposed modified geometry. L3 = 03.40, L4 = 4.23, L5 = 9.30, L6 = 7.10, L7 = 6.3, L8 = 6.2, L9 = 3, L10 = 4.75, Lg = 10.25.

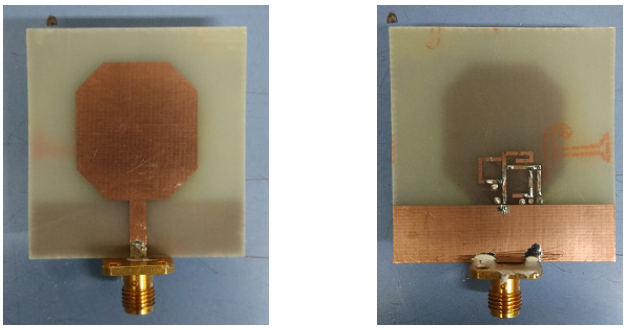


Figure 8. Photograph of the modified geometry in Figure 7.

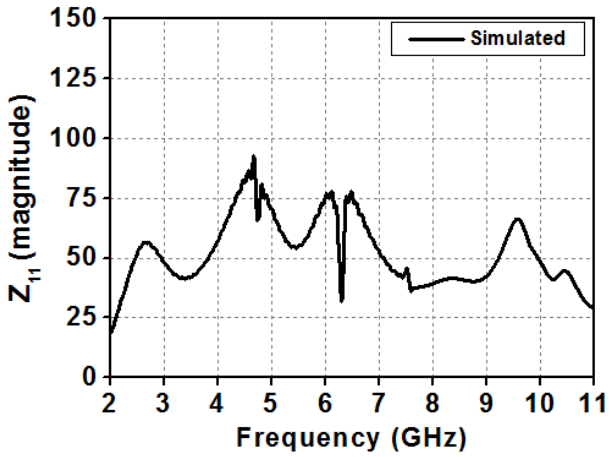
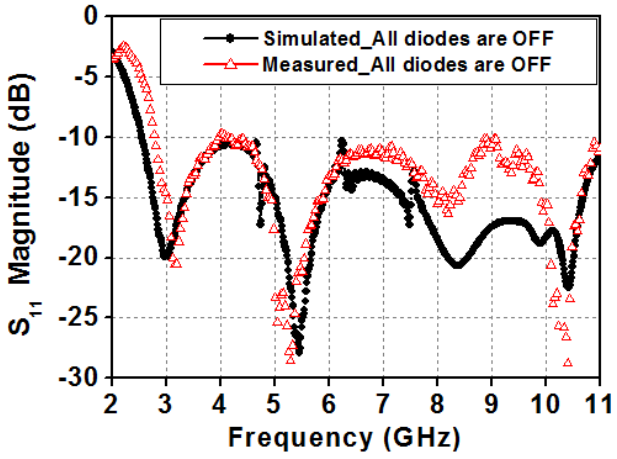
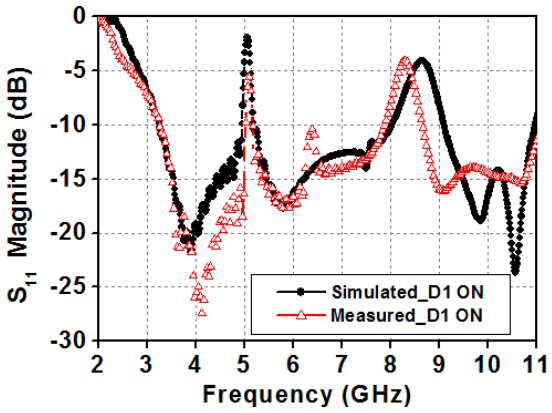


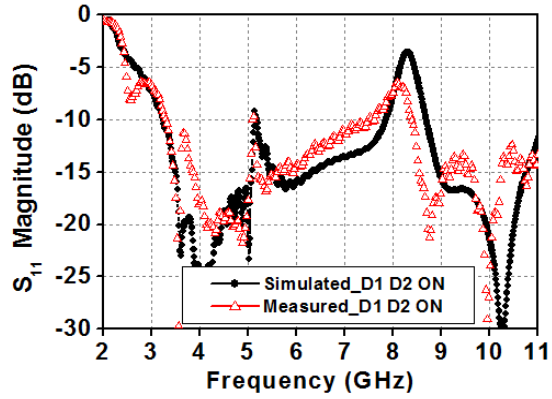
Figure 9. Input impedance of the proposed antenna when all diodes are OFF.



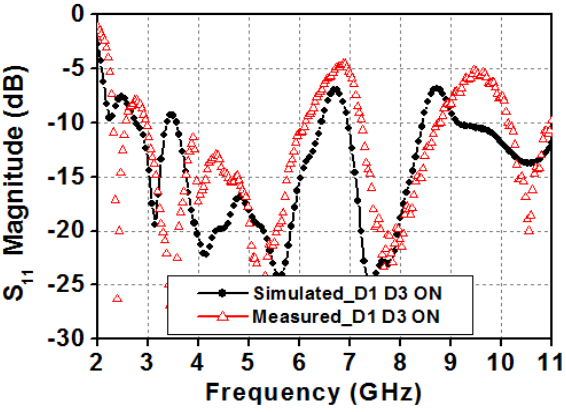
(a)



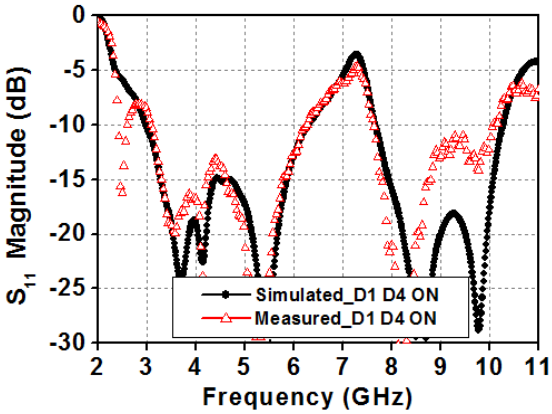
(b)



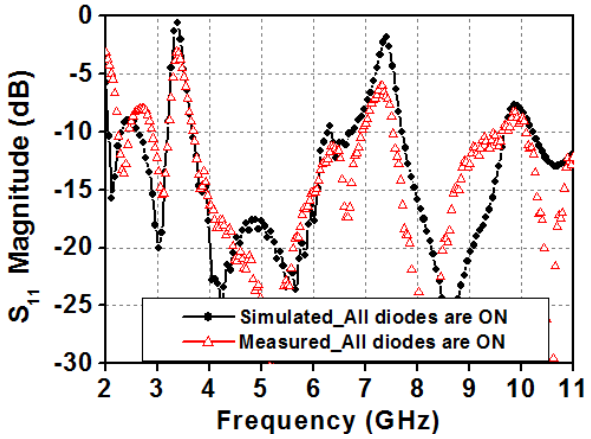
(c)



(d)



(e)



(f)

Figure 10. Reflection coefficients of the modified antenna as shown in Figure 7.

4. Switchable and tunable antenna design

The switchable and tunable reconfigurable UWB modified octagonal monopole patch antenna is proposed by modifying the structure as shown in Figure 11. Front side of the structure is same as used in the previous one (Section 2).

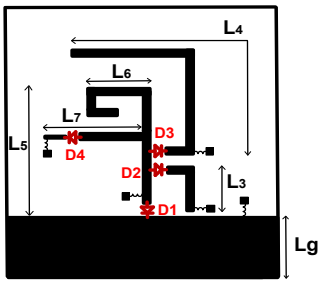


Figure 11. Switchable and tunable notch-band antenna: L3 = 5.1, L4 = 14.8, L5 = 9.45, L6 = 5.5, L7 = 11.2, Lg = 10.3.

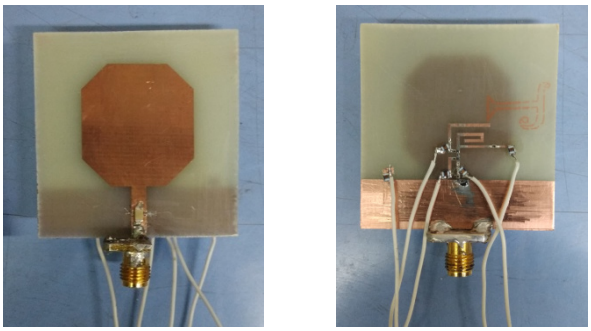


Figure 12. Photograph of the switchable and tunable notch-band antenna as given in Figure 11.

One varactor and three PIN diodes are used to make antenna switchable and tuneable. Photograph of the printed circuit is shown in Figure 12. An input impedance plot, with all diodes OFF, is shown in Figure 13. In the first state, when all diodes are OFF, the reflection coefficient is below -10 dB from 2.94 to 11.85 GHz as shown in Figure 14(a). In state 2, diode D1 is ON, two notched-bands are created at 4.8 and 8.25 GHz as shown in Figure 14(b). In state 3, diodes D1 and D2 are ON, the notched-band created at 8.25 GHz is suppressed by connecting L3 stub as shown in Figure 14(c). In state 4, diodes D1 and D3 ON, notched-bands are shifted

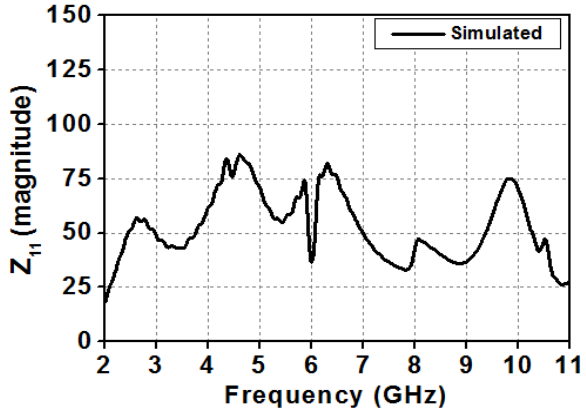
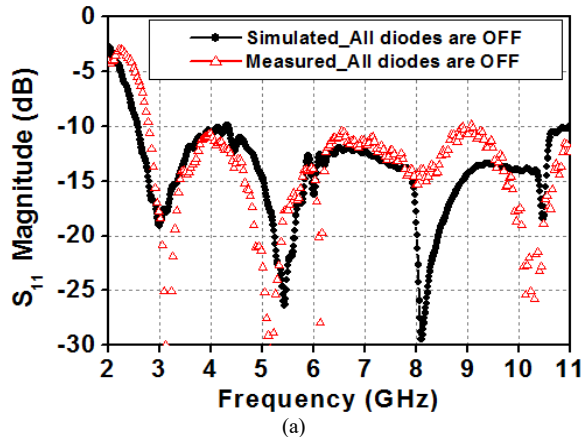
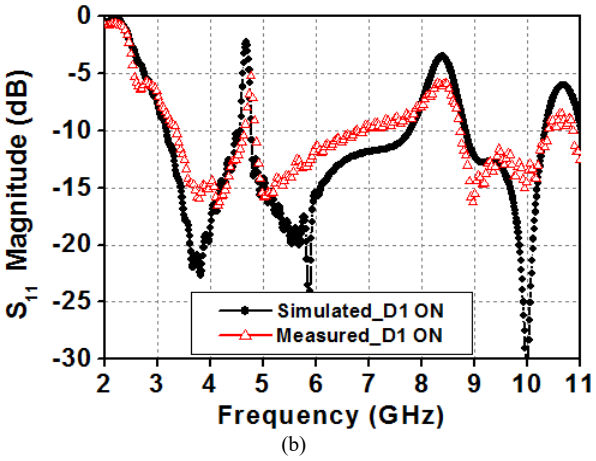


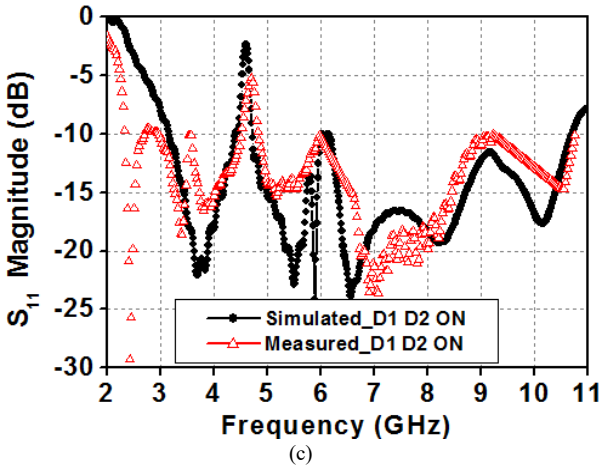
Figure 13. Input impedance of the proposed antenna when all diodes are OFF.



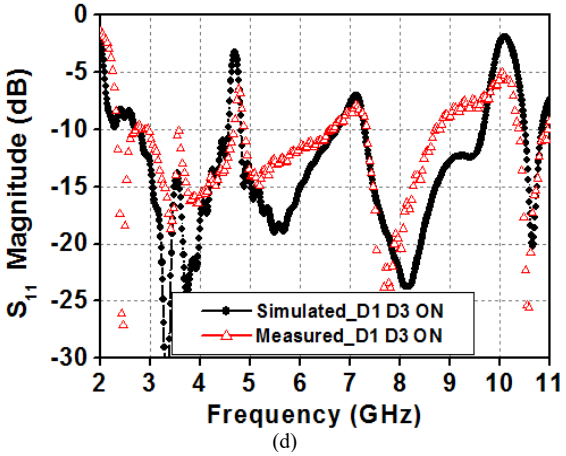
(a)



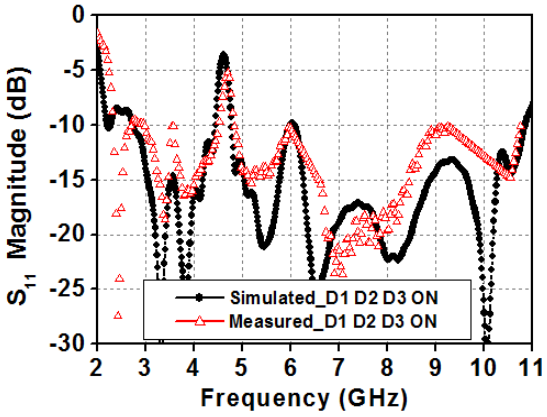
(b)



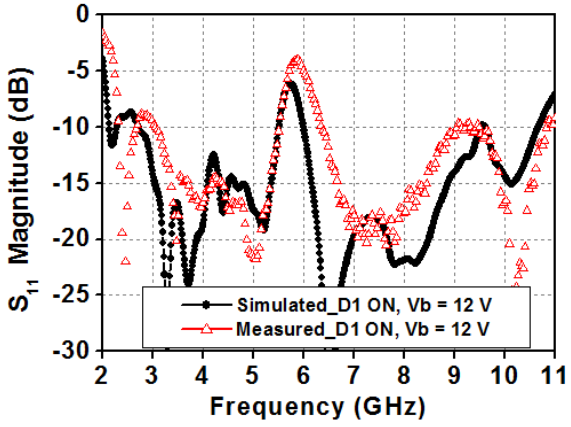
(c)



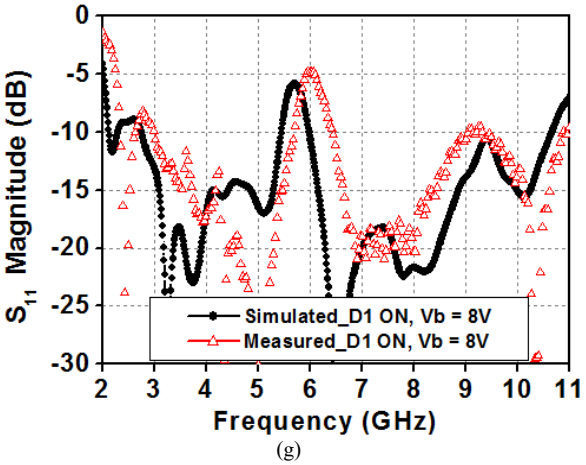
(d)



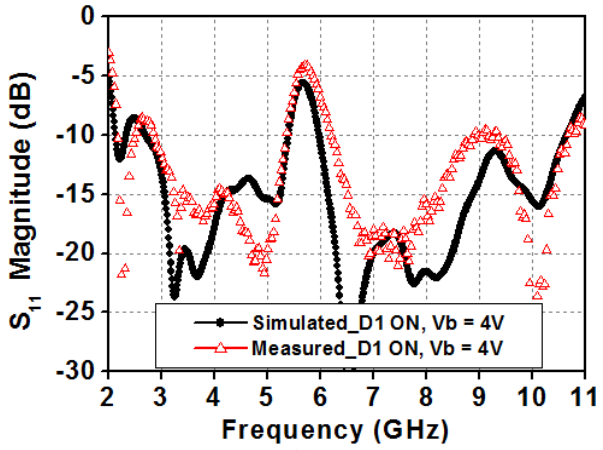
(e)



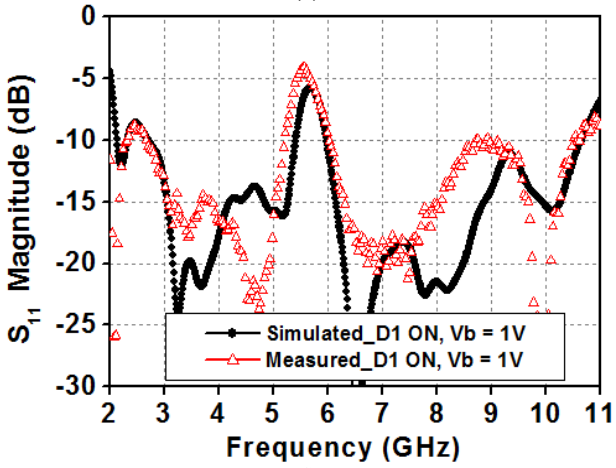
(f)



(g)



(h)



(i)

Figure 14. Reflection coefficients of the modified switchable and tunable antenna.

by changing the overall reactance as shown in Figure 14(d). In state 5, diodes D1, D2 and D3 are ON, one notch is obtained at 4.75 GHz as shown in Figure 14(e).

Continuous tuning is also obtained from the structure, diode D4 (varactor diode, model MA46H120) is integrated as shown in Figure 11. Capacitance of the diode varies from 0.1 pF to 1 pF with bias voltage varies from 12 V to 1V, respectively. The results are plotted with different bias voltages as shown in Figure 14(f) to (i). Notch-bands are shifted from higher to lower side with centre frequencies of 6 and 5.5 GHz, respectively.

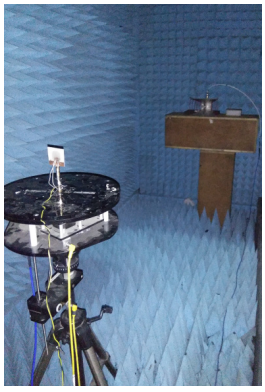
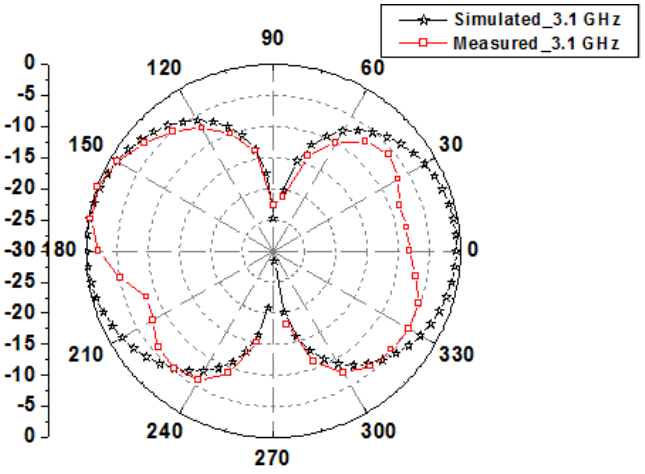
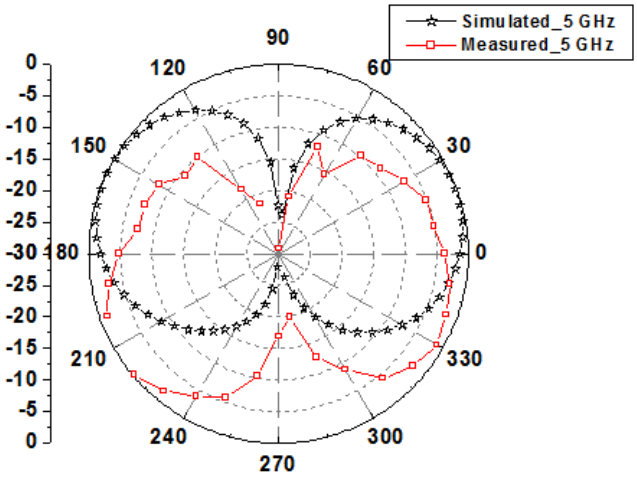


Figure 15. Photograph of the measurement setup.

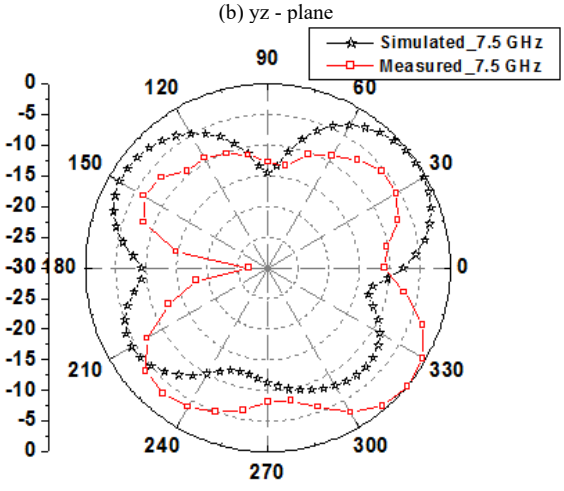
The radiation patterns of the switchable and tuneable octagonal monopole patch antenna are measured in the anechoic chamber. Photograph of the measurement setup is shown in Figure 15. A double-ridged horn antenna is used to illuminate the test antenna. The proposed (test) antenna is placed on a rotating table which rotates horizontally. Figure 16 shows the normalized radiation patterns. The cross-polarization level is better than 12 dB at all measured frequencies.



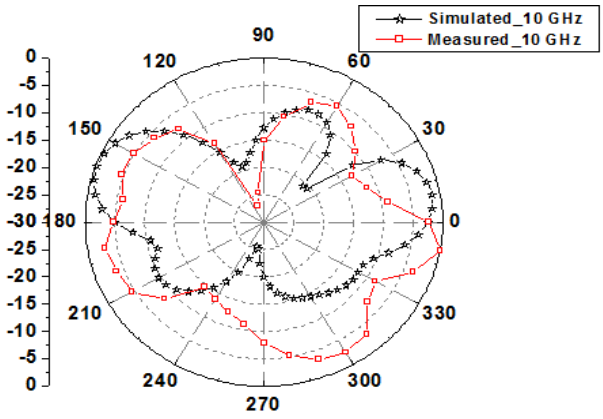
(a) yz - plane



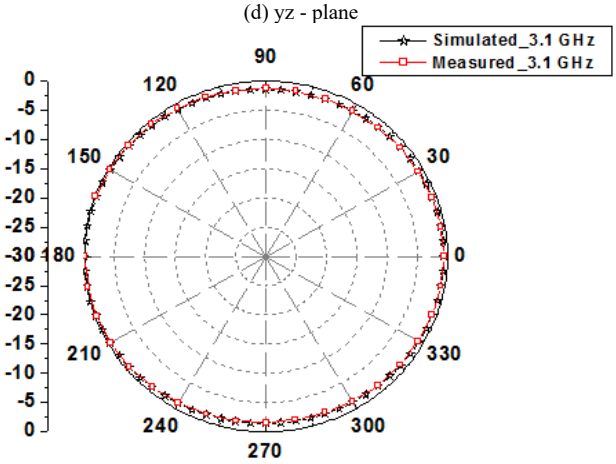
(b) yz - plane



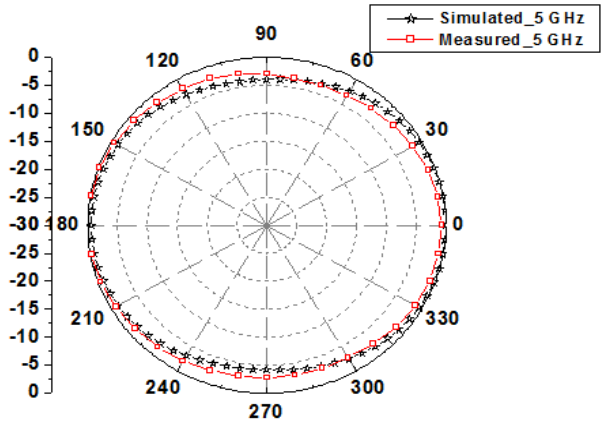
(c) yz - plane



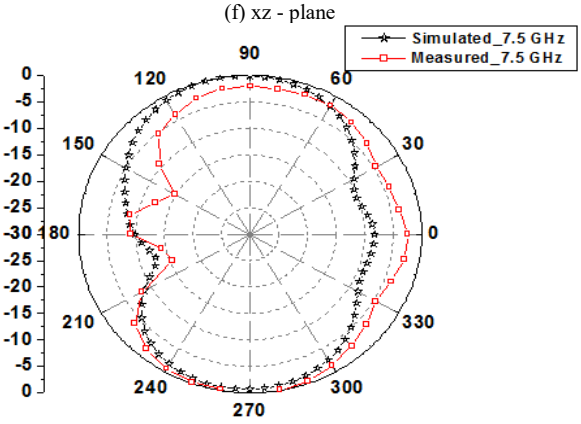
(d) yz - plane



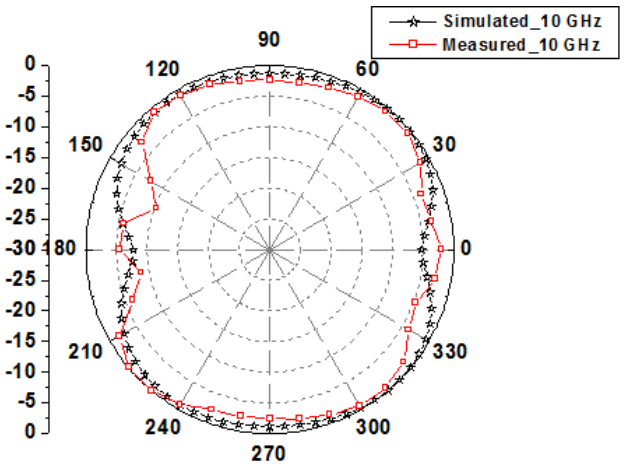
(e) xz - plane



(f) xz - plane



(g) xz - plane



(h) xz - plane

Figure 16. Normalized radiation patterns of the proposed switchable and tunable antenna.

The measured gains are better than 3.12 dB in all states except at the notched-bands. The gains in the notched-bands are below -7.4 dB. The discrepancies in the results are due to fabrication imperfections and use of insulated wires. Insulated wires affect the radiation pattern. Further, the sharp notches can be obtained from this structure by using lumped elements. The filters made from microstrip line have some bandwidth which is more than that of achieved with the lumped elements.

Table 2: Comparisons of the proposed work with earlier reported ones

	Frequency Range (GHz)	No. of notched-bands	Total size L×W (mm^2)	Type of switching
[9]	2 – 12.5	Two notched-bands @ 5.3, 7.4 GHz	38.5 × 46	Discrete
[17]	3.1 – 14	One tuning band @ 5.1-5.9	22 × 22	Continuous
[18]	3.1 – 12*	One tuning band @ 3.6-5.2	18 × 21	Continuous
[19]	2.9 – 12	Four notched-bands @ 3.5, 4.6, 5.2, 5.8 GHz	30 × 31	Discrete
[20]	3.1 – 10.7	One notched-band @ 5 GHz	30 × 27.4	Discrete
This work (with varactor diode)	2.8 – 11.5	Five discrete notched-bands @ 4.5, 4.75, 4.85, 8.25, 10 GHz and one tuning band @ 5.5 – 6 GHz	42 × 40	Discrete & Continuous

* Estimated from the data given in the paper

Comparison with earlier reported papers is given in Table 2. In [9], one quasi-complementary split-ring resonator is etched to achieve dual notched frequency bands while in [20], one notched-band is created by embedding a π-shaped slot in the structure, but the advantage is the compact structure. In [19], four notched-bands are obtained by inserting four U-shaped and one L-shaped slot. Continuous switching was achieved in [17] and [18]. In all these designs except [17], notched-bands were created by perturbing the radiating structure. In this paper, the notched-bands are created by using stubs as band stop filters (BSFs) on the back side of a radiator instead of changing the front side which has a radiator and feed line. In such cases, the radiation patterns are less affected while changing the different states of the antenna [17]. In electromagnetic scattering environment, antenna must have freedom to suppress the selected band. The proposed antenna can suppress the bands with centre frequencies of 4.5, 4.75, 4.85, 8.25, 10 and 5.5-6 GHz.

5. Conclusion

An experimental investigation of reconfigurable UWB modified octagonal microstrip monopole patch antenna with switchable and tunable band-notched characteristic is proposed in this paper. Initially, the design is started with creation of switchable notched-bands, after that some additional notches are obtained by using a modified structure. Finally, by employing a varactor diode, the tuning of notched-bands is achieved. The proposed designs are simple in geometry. The proposed antenna has freedom to suppress many interfering bands. The tunable band-notched reconfigurable UWB antenna is suitable for modern wireless communication systems.

References

[1] T. Kaiser, F. Zheng, and E. Dimitrov, “An overview of ultra-wide-band systems with MIMO,” Proc. IEEE, 97(2), 285–312, Feb. 2009.

[2] J. Ren, W. Hu, Y. Z. Yin, and R. Fan, “Compact printed MIMO antenna for UWB applications,” IEEE Antennas Wireless Propag. Lett., 13, 1517–1520, 2014.

[3] Wang-Sang Lee, Dong-Zo Kim, Ki-Jin Kim and Jong-Won Yu, "Wideband planar monopole antennas with dual band-notched characteristics," IEEE Trans. Microw. Theory Tech., 54(6), 2800-2806, June 2006.

[4] P. Gao et al., “Compact printed UWB diversity slot antenna with 5.5-GHz band-notched characteristics,” IEEE Antennas Wireless Propag. Lett., 13, 376–379, 2014.

[5] J. -M. Lee, K. -B. Kim, H. -K. Ryu, and J. -M. Woo, “A compact ultra wideband MIMO antenna with WLAN band-rejected operation for mobile devices,” IEEE Antennas Wireless Propag. Lett., 11, 990–993, 2012.

[6] L. Kang, H. Li, X. Wang and X. Shi, "Compact Offset Microstrip-Fed MIMO Antenna for Band-Notched UWB Applications," IEEE Antennas Wireless Propag. Lett., 14, 1754-1757, 2015.

[7] T. K. Roshna, U. Deepak, V. R. Sajitha, K. Vasudevan and P. Mohanan, "A Compact UWB MIMO Antenna With Reflector to Enhance Isolation," IEEE Trans. Antennas Propag., 63(4), 1873-1877, April 2015.

[8] J.-F. Li, Q.-X. Chu, Z.-H. Li, and X.-X. Xia, “Compact dual band notched UWB MIMO antenna with high isolation,” IEEE Trans. Antennas Propag., 61(9), 4759–4766, Sep. 2013.

[9] W. T. Li, Y. Q. Hei, W. Feng and X. W. Shi, "Planar Antenna for 3G/Bluetooth/WiMAX and UWB Applications With Dual Band-Notched Characteristics," IEEE Antennas Wireless Propag. Lett., 11, 61-64, 2012.

[10] S. Wei, Y. Yin, Y. Huang, S. Fan, L. Kang and X. Li, "Novel smiling face-shaped antenna for dual band-notched ultra-wideband applications," 2010 IEEE International Conference on Ultra-Wideband, Nanjing, 2010, 1-3.

[11] He Huang, Ying Liu and Shuxi Gong, “A novel uniplanar differentially-fed UWB polarization diversity antenna with dual notch bands,” 2015 9th European Conference on Antennas and Propagation (EuCAP), Lisbon, 2015, 1-4.

[12] D. Yu, G. Xie, Z. Liao and W. Zhai, "Novel dual band-notched omni-directional UWB antenna on double substrates crossing," 2011 4th IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, Beijing, 2011, 95-97.

[13] L. Kurra, M. P. Abegaonkar, A. Basu and S. K. Koul, "A band-notched UWB antenna using uni-planar EBG structure," 2013 7th European Conference on Antennas and Propagation (EuCAP), Gothenburg, 2013, 2466-2469.
[14] X. Wang, Z. F. Yao, Z. Cui, L. Luo and S. X. Zhang, "CPW-fed band-notched monopole antenna for UWB applications," 2008 8th International Symposium on Antennas, Propagation and EM Theory, Kunming, 2008, 204-206.
[15] J. Dong, G. Fu, J. Zhao and X. Wang, "Dual band-notched UWB antenna with folded SIRs," 2012 International Conference on Microwave and Millimeter Wave Technology (ICMMT), Shenzhen, 2012, 1-3.
[16] A. K. Tagantsev, V. O. Sherman, K. F. Astafiev, J. Venkatesh, and N. Setter, "Ferroelectric Materials for Microwave Tunable Applications," Journal of Electroceramics, 11, 5-66, 2003.
[17] R. Zaker, C. Ghobadi and J. Nourinia, "Novel Modified UWB Planar Monopole Antenna With Variable Frequency Band-Notch Function," IEEE Antennas Wireless Propag. Lett., 7, 112-114, 2008.
[18] Mahdi Nouri and Sajjad Abazari Aghdam, "reconfigurable uwb antenna with electrically control for triple on-demand rejection bandwidth," Microwave Optical Technology Lett., 57(8), August 2015.
[19] Z. H. Wu, F. Wei, X. W. Shi and W. T. Li, "A compact quad band-notched UWB monopole antenna loaded one lateral L-shaped slot," Progress in Electromagnetics Research, 139, 303–315, 2013.
[20] Y. -L. Zhao, Y. -C. Jiao, G. Zhao, L. Zhang, Y. Song and Z. -B. Wong, "Compact planar monopole UWB antenna with band-notched characteristic," Microwave Optical Technology Lett., 50(10), 2656–2658, Oct. 2008.
[21] Ruchi Paliwal, Rajesh K. Singh and Shiban K. Koul, "Reconfigurable UWB Monopole Antenna with Switchable Frequency Notched Bands," IEEE Applied Electromagnetics Conference, Dec. 2017, 1-4.
[22] CST Studio Suite, [Online]. Available: http://www.cst.com/
[23] Data Sheet of MA4SPS402 PIN diodes, MA-Com, Application Note.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 211-217 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Psychoacoustical Approach in Soundscape Characterization

Mia Suhanek*, Ivan Djurek, Sanja Grubeša, Antonio Petošić

University of Zagreb, Faculty of EE and Computing, Department of Electroacoustics, Unska 3, HR-10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received: 21 June, 2018
Accepted: 28 July, 2018
Online: 31 July, 2018

Keywords:
sound environment
soundscape
loudness
noise

ABSTRACT

Four different soundscapes with different loudness and spectral distributions were recorded in Zagreb (capital of Croatia) and reproduced to two groups of listeners while they were performing a concentration demanding task – a simplified variation of the memory game. These two groups of test subjects listened to the same soundscapes but with different loudness distribution. The game results versus time were recorded, analyzed and compared to questionnaire answers, along with loudness and property changes of each soundscape. The goal was to determine which of the sound events causes the largest distraction and to establish a connection between these results and the annoyance level of each soundscape. As expected, the loudest soundscape resulted in the lowest score and caused the largest annoyance. However, other factors proved to cause listeners' distraction, mainly unexpected sounds that differ from the main sound print. In that sense, this paper also deals with a concept of creating enjoyable sound environment in urban places.

1. Introduction

The initial concept of soundscape was proposed as an attempt to create an analytical perspective that would explain the total acoustic environment over time and across cultures [1]. Schafer wrote in 1977: "Ecology is the study of the relationship between living organisms and their environment. Acoustic ecology is the study of sounds in relationship to life and society." He also suggested three different types of noise: 1) unwanted sound, 2) unmusical sound (defined as non-periodic vibration and 3) any loud sound and disturbance in signaling systems. In his book he deals with the idea that every city has a unique urban soundscape and furthermore, every part of the city has a unique soundscape. An example for that is a siren in a noisy environment can be barely noticeable, while the same siren in a different but calmer neighborhood can result in a temporal or chronic hearing loss if the attack is sudden [2],[3]. The soundscape studies are multidisciplinary and include: acoustics, psychoacoustics, otology (study and treatment of the ear) and noise reduction. Most soundscape studies concern the qualitative analysis of soundscapes; however, the methods for evaluating a soundscape vary depending on the purposes of the studies and the researchers conducting them [4],[5]. So far soundscape research has been oriented mainly on evaluating a soundscape in a qualitative way, e.g. assessing the soundscape pleasantness or rather its unpleasantness [6]-[9]. In order to describe a certain soundscape in greater detail and provide a more comprehensive approach to soundscape research, it is necessary to use other methods which could also define soundscape characteristics quantitatively. The basis for acoustical characterization can be achieved by using a set of subjective descriptors, for example bipolar adjectives, that describe specific soundscape characteristics such as loudness or a level of distraction or annoyance by a certain sound and are then assigned a numerical score which enables and facilitates the statistical analysis of specific sound properties [2],[10]-[12].

From a quantitative point of view, statistical relevance is of utmost importance in efforts to describe or quantify a listener's perception and evaluation of a soundscape. Taking into account several variables such as loudness, pitch strength and fluctuation of pitch strength among others, it is possible to calculate the annoyance level of a certain sound or a soundscape [13]-[16]. Another part of soundscape and sound analysis deals with their distraction properties and the way in which characteristic sounds influence people performing logical, mathematical and other concentration demanding tasks [10],[11].

People are exposed to different sound environments on a daily basis. Long exposure to a specific sound environment results in eventual adaptation to that environment, and small and expected loudness changes in soundscape do not significantly influence one's perception of that soundscape [2]. Frequent exposure to even a slight change in our everyday sound environment would also result in adaptation; furthermore, providing the loudness levels were not significantly higher, these changes would not be perceived as annoying [3].

*Mia Suhanek, University of Zagreb, Croatia, 00385 1 6129840,
Email: mia.suhanek@fer.hr

In this paper, we aimed at analyzing and comparing in what way certain soundscapes with different loudness distributions distract listeners performing a rather simple but concentration-demanding task, like a simplified variation of the memory game. Since the test subjects received no training prior to undergoing the testing and were only given technical information pertaining to the testing environment, this game was chosen due to the familiarity of the test subjects with its format. Furthermore, the level of concentration required for this simplified variation of the game corresponds to day-to-day situations where people perform basic tasks while being exposed to different sounds which do not require their focus, but only cause distraction. We wanted to monitor listeners' reactions to different sounds in the background of the main theme of a soundscape and to determine whether these sound events cause the distraction on the same level as louder, familiar and expected sounds of a certain soundscape.

In general, we intentionally opted for different spectral content, loudness distribution, duration and rate of change of individual louder events in the recorded soundscapes. The underlying idea was that the listeners perform a certain task for the duration of all soundscapes. We wanted for the listeners to perceive the soundscapes subconsciously, rather than actually listen to the recordings. The task involved playing a game, thus requiring a higher level of concentration. The results of the game for all listeners were compared to questionnaire responses which lead us to conclude that, when analyzing soundscape in general, listeners base the level of their annoyance mainly on loudness. On the other hand, distraction can also be attributed to unexpected sounds that are somewhat different from the main sound print.

Furthermore, a soundscape by definition includes a certain environment and therefore a large number of sound sources that affect human activity, concentration and mood which is why soundscape research investigates the conditions and reasons of its origin as well as the level of human soundscape perception and evaluation[1].

Tim Beatley also explains in his paper "Celebrating the Natural Soundscapes of Cities" (2013) the importance of soundscape of the city in a way that the city should be enriched with natural sounds. There are also other studies that have shown that people especially enjoy the sound of water[6]. In that sense, we can conclude that the real challenge is how to create an enjoyable soundscape in nowadays overpopulated cities.

Unfortunately, the main concern of the society in general, is still only noise management and legislative. We can conclude that nowadays, urban planners have a big challenge before them. While fitting visual aesthetics with solutions for transport, waste and energy, they need to bear in mind the natural soundscape preservation perspectives.

2. Urban Soundscapes in Zagreb

The first recorded soundscape was a children's park situated in the western suburb of Zagreb and surrounded by large housing blocks. Figure 1 shows a spectrogram of the recorded children's park environment excerpt. The spectrogram is relatively wide stretching above 10 kHz, with very short leaps in loudness change which are up to 20 dB higher when compared to the steady part of the recording.

The second recorded soundscape was the expressway stretching from the east to the western exit of the city. Traffic in this avenue is almost always heavy and dense so the soundscape included sounds coming from cars, buses and trucks passing by and an audible traffic signal for the visually impaired (see Figure 2). The frequency spectrum is narrower compared to the children's park environment and is mainly concentrated at frequencies below 1 kHz. Sudden and large changes in loudness were not as unexpected for the expressway environment as they were in the case of the children's park.

The third soundscape was a stream on the outskirts of western part of Zagreb, in a small forest surrounded with local roads. The recording included sounds coming from the stream, nearby traffic, children playing and a small dog barking. Figure 3 shows the frequency spectrum of this recording excerpt. It is narrow and concentrated in the low frequencies making the sudden loudness changes (dog barking, children screaming etc.) very discernible.

The fourth soundscape was an industrial hall which produces rail vehicles situated in the eastern part of Zagreb. Figure 4 is a spectrogram of this soundscape which included the sounds of power tools such as grinders and drills, the sound of a hammer hitting metal and music coming from the radio. These specific sounds are very loud, as shown on the spectrogram (see Figure 4). The frequency spectrum is up to 10 kHz wide, with short loudness changes rising to 30 dB above the average level.

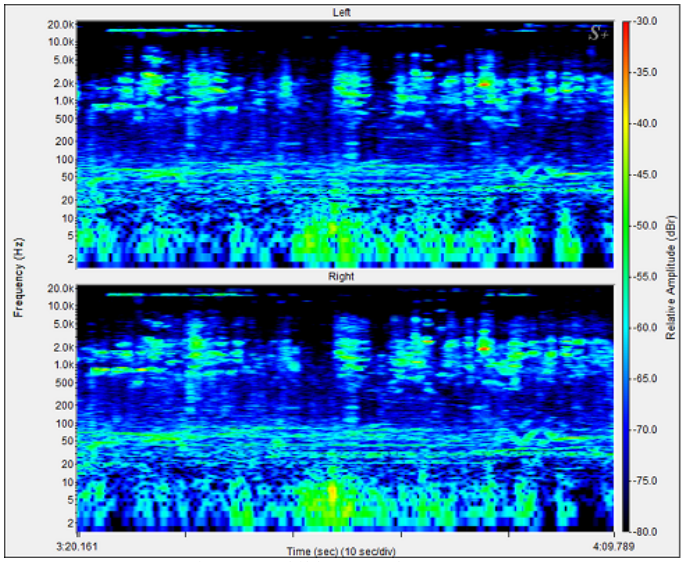


Figure 1. Children's park spectrogram

These four soundscapes were chosen due to their relatively different sound characteristics. An average person living in an urban environment is familiar with three of them: the children's park, the expressway and the stream soundscape. However, the industrial hall soundscape is less common, to say the least. For the research, comparing familiar with unfamiliar soundscapes enabled us to analyze whether the former was perceived less distracting than the latter. The research participants were divided into two groups: the control (marked *CG*) and experimental group (marked *EG*). The experimental group listened to the soundscapes with frequent sudden and short loudness changes, and the control group listened to the same soundscape but with lower loudness changes. Loudness versus time diagram were also created for a specific soundscape and listening group (see Figures 5a, 5b, 5c and 5d)

based on Figures 1 to 4. In Figures 5a, 5b, 5c and 5d the soundscape recording for the experimental group is marked with a solid line while the recording for the control group is marked with a dotted line.

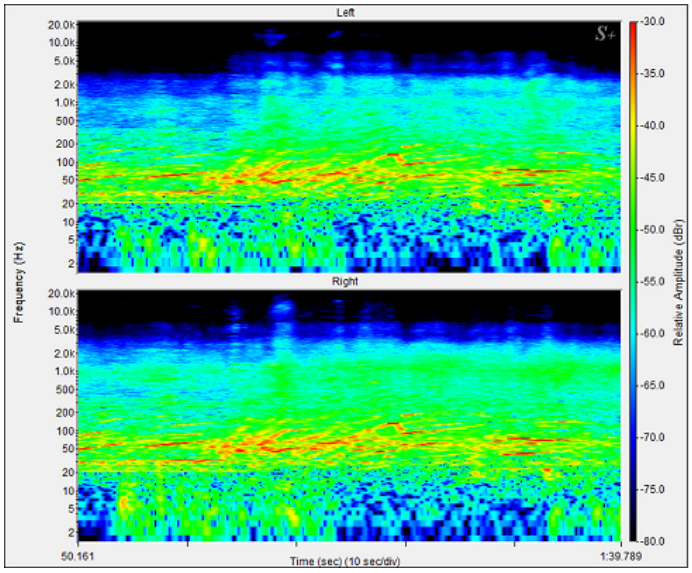


Figure 2. Expressway spectrogram

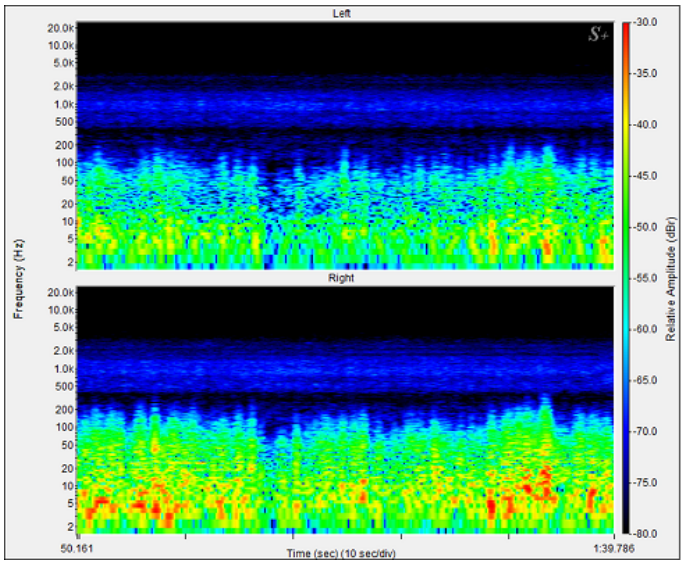


Figure 3. Forest stream spectrogram

3. Listening Tests

Four characteristic urban soundscapes were recorded using the *soundwalk* method [5,17]. The soundscape samples were recorded using the M-AUDIO recorder [18] and a pair of binaural microphones, with a 48 kHz sampling rate and a 16 bit quantization standard. The *soundwalks* were performed at different times of the day, different days of the week, always in nice, dry and sunny weather in March 2018. These recordings were performed at the *soundwalker's* height so that the patterns obtained would be as similar to the natural binaural listening of people residing in these soundscapes. The recordings were then reproduced to two groups of listeners: the control group (CG) and the experimental group (EG). The experimental group listened to the soundscapes with frequent sudden and short loudness changes, and the control group listened to the same soundscape but with lower loudness changes.

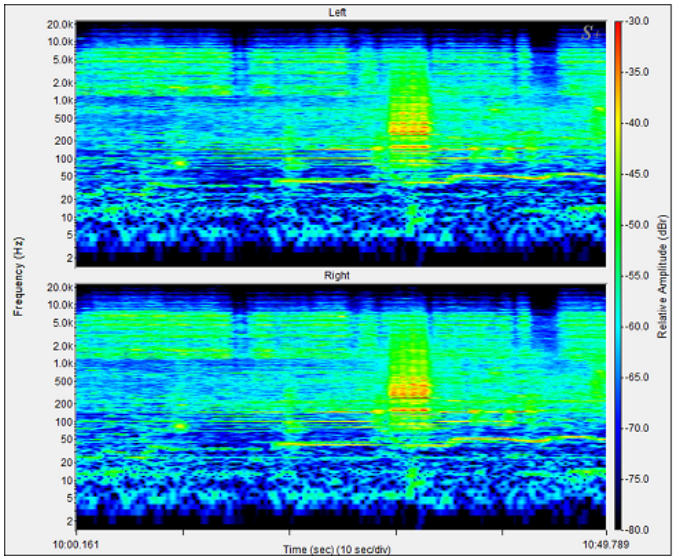


Figure 4. Industrial hall spectrogram

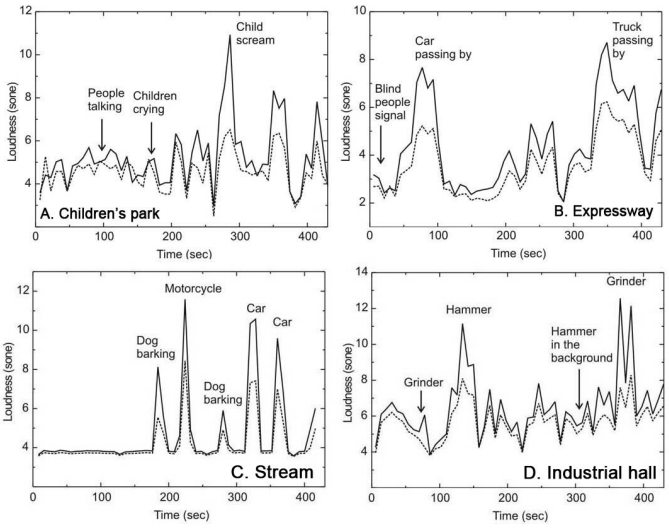


Figure 5. Average loudness versus time diagram for (the control group is presented with a dotted line and experimental group with a solid line): a) children's park, b) expressway, c) stream and d) industrial hall

The recordings were reproduced using AKG K55 closed electrodynamic headphones with an average sound pressure level of 50 dB(A) in the steady part of the recorded soundscape. Loudness was calculated using the established Zwicker method [13],[14]. Free field equalization was used. It is important to emphasize that average sound pressure level of 50 dBA refers to the recordings listened by the experimental group while the recordings for the control group are corrected by a compressor with a limit value of 60 dBA and a compression ratio of 3:1. The listeners' groups had equal female-to-male ratio, with median age of 24. Each of the listeners listened to all four recordings with a time gap of at least one week in order to facilitate recovery time and eliminate any potential influence of one recording on another. The listeners were not informed as to the content of the recording. The order of the listening of the soundscapes was fixed for all listeners which is in line with the established psychological and statistical research methodology and praxis [19],[20], and is as

follows: the children's park soundscape, the expressway soundscape, the industrial hall soundscape and the stream soundscape. The results wouldn't have been different if we chose a different order of the soundscape listening. However, once you have chosen the order of the listening it has to remain the same for all participants of the study. For example: if a participant is placed in a control group (he or she is not of course aware of it) then he or she has to stay in the control group till the end of the study. Furthermore, if one participant listened the soundscapes in a certain order then the rest of the participants must have the same order due to psychological methodology and its requirement that all participants have to have the same testing conditions. Thus, this way of research enables the researcher to have a better control over the study.

Table 1 Most annoying sounds for the experimental and control groups (EG- experimental group, CG- control group)

Children's park	**EG**	**CG**
children screaming	44,8%	34,7%
sound of the swing	27,6%	20,4%
children crying	10,4%	8,2%
kindergarten teacher	13,8%	16,3%
all of the above	1,7%	8,2%
nothing bothered me	1,7%	12,2%

Expressway	**EG**	**CG**
vehicle horn	13,3%	2,4%
audible traffic signal for the visually impaired	11,1%	17,1%
sound of cars	21,1%	43,9%
all of the above	0,0%	7,3%
nothing bothered me	24,5%	29,3%

Industrial hall	**EG**	**CG**
sound of power grinders	18,2%	41,9%
sound of hammers	70,4%	38,7%
music from the radio	0,0%	1,6%
all of the above	2,3%	1,6%
nothing bothered me	9,1%	16,2%

Stream	**EG**	**CG**
small dog barking	27,6%	5,2%
children screaming	20,7%	10,7%
bus passing by	37,9%	7,3%
stream bubbling	0,0%	10,5%
all of the above	13,8%	0,0%
nothing bothered me	0,0%	66,3%

Additionally, a questionnaire was composed in order to get a more detailed description of which particular sounds in a given soundscape were the most annoying and why. The listeners had to provide an answer to a direct question: Which of the given sounds in the soundscape bothered you the most? This form of research enabled us to establish a connection between annoying sounds and the results of the game (e.g. listeners' distraction). For every soundscape we have supplied a list of potentially most annoying sounds. However, the list was far from selective as it contained all the characteristic sound events appearing in a certain soundscape. Our intention was to provide a reminder for the listeners considering the long duration of the soundscape samples, rather than to influence their answers. These results are laid out in Table 1. In addition to the list of the characteristic sound events, the questionnaire provided the "all or nothing of the above" answer options for each soundscape.

4. The Game

During the first listening of each soundscape, the listeners had to solve an interactive game, similar to the traditional children's memory game, programmed in MATLAB (see Figure 6).

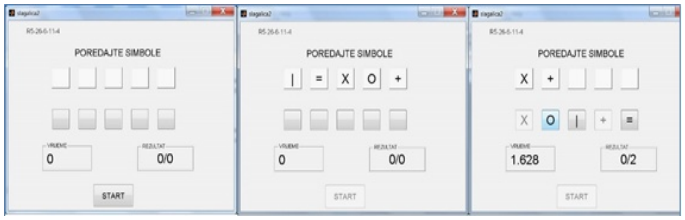


Figure 6. The game interface in MATLAB

The listeners played the game on the computer, while at the same time listening to a certain soundscape. The game uses five cards with simple mathematical symbols (X, O, I, +, =). The symbols on the cards are revealed to the listener in a random order for four seconds. The goal is for the listener to line up the symbols on the cards according to the order in which they appeared. Time allotted for this task is seven seconds, with a two second pause before the next layout of the cards. If the listener makes a mistake while arranging the symbols on cards, a new hand of cards is drawn automatically. The game is not intended to be difficult or complicated, but rather to serve as means of assessing listener's concentration or the absence thereof during specific sound events in each soundscape, i.e., sudden loudness changes. For the duration of each soundscape, the program generates orders of symbols and records the listener's score. In this way we could establish and analyze the potential correlation between the results of the game and time, as well as identify certain parts of a soundscape that could have caused the distraction occurring at a specific point in time. The results of the game, representing the success ratio of each move, are generated numerically. For instance, if a listener successfully arranged all the cards in the given time frame, their result for this move equaled 1. If a listener successfully arranged only 3 out of 5 cards, the result was 3/5 = 0.6. We then calculated and compared the average scores for each soundscape overall.

Finally, for each listener, we created an average score versus time diagram for a specific soundscape and listening group. All of these were then used in calculating the overall average score

versus time diagram. Figure 7a, 7b, 7c to 7d show these diagrams for all four soundscapes.

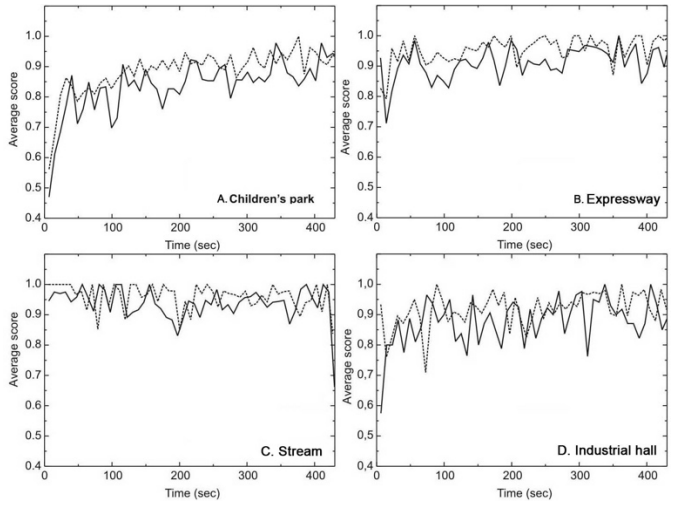


Figure 7. Average score versus time diagram for (the control group is presented with a dotted line and experimental group with a solid line): a) children's park, b) expressway, c) stream and d) industrial hall

5. Research Results

The questionnaire responses describing the listeners' assessment of the soundscapes showed that, in the majority of cases, annoyance can be assigned to loud sounds - mainly those louder than the main theme of the soundscape. Furthermore, a comparison between the experimental and control group soundscapes shows that, in general, loudness has the strongest influence on the annoyance factor. On the other hand, a more detailed analysis shows that significant loudness changes are not the only factor contributing to listeners' distraction.

It is interesting to notice that when comparing two relatively loud soundscapes - the children's park and industrial hall soundscapes, the percentage of listeners not bothered with any of the characteristic sound events was higher in the case of the loudest soundscape with artificial sounds – the industrial hall. Based on the content, the comparison of these two soundscapes shows that the louder sounds are much more varied in the former soundscape. Even though the industrial hall soundscape features a greater number of louder sounds, they are quite similar and consist mainly of the sound of power tools. In the case of the children's park soundscape, individual sound events are much more varied in terms of frequency content, loudness distribution and rate of change.

In the case of the expressway soundscape, a larger percentage of listeners stated that nothing bothered them. This soundscape is in general quieter, but also consist mainly of the same sounds, with very narrow loudness distributions. In addition, fewer loud sounds differ from the main soundscape print in a way to annoy and distract the listeners.

The specific nature of the stream soundscape reflects in the longest period of the steadiness of the main theme. However, all the listeners in the experimental group reported some type of annoyance. Even though this soundscape features less sudden loudness changes than the previous three, these transitions are considerable in comparison with the steady part of the recording. Also, a large number of listeners in the control group stated that nothing bothered them which is in compliance with various studies in this field in terms of conclusion that people prefer the sound of water.

Upon careful examination of the game results, all the diagrams show a similar pattern (Figure7a, 7b, 7c and 7d). The average move score tends to slightly increase toward the end of the soundscapes. This effect is most evident in the children's park soundscape. Moreover, the first moves for all listeners are most inaccurate, while the average scores tend to increase toward the end of the soundscape. The reasons for this are twofold: adaptation to the game environment and adaptation to the soundscape. Despite the fact that all the listeners played some type of a memory game prior to this testing and were therefore familiar with the pattern of the game, their initial moves were considerably less accurate than those made later in the game. The recorded average move time stabilized after the first few moves, and this can be attributed both to the adaptation to the game environment and adaptation to the soundscape factors. This effect is less obvious for soundscapes with smaller loudness change and for the control group soundscapes where loudness changes were intentionally lower, which is indicative of the importance of the occurrence of sudden and unexpected sounds.

Comparing the average score for all four soundscapes (see Table 2), the children's park soundscape has the lowest average score, immediately followed by the industrial hall soundscape, expressway and stream soundscapes. There is an obvious difference for the experimental and control group soundscapes. As mentioned, soundscapes with lower loudness changes have lower distraction properties than those with higher loudness changes. This corresponds to the questionnaire results, where larger percentage of listeners of the control group soundscapes stated that nothing bothered them.

If we compare the average score with loudness diagrams for a specific soundscape, in some cases the lower score did not correspond to higher loudness change, as would be expected. Several other factors that affected a lower game score must be taken into consideration. For example, the time segment around the 100th second in the children's park soundscape for the experimental group shows a lower average score even though there are no considerable loudness changes in this part of the soundscape. However, this interval features a short conversation which immediately distracted the majority of listeners. The same distraction pattern occurred whenever an indistinct conversation could be overheard in this soundscape. It seems that larger loudness changes in the form of children screaming did not distract the listeners playing the game as would be expected.

A similar example can be found in the industrial hall soundscape for the experimental group. Around the 300th second, a sound of a hammer could be heard in addition to that of power grinders. This sound stimulus proved to be unexpected and very distracting despite the fact that the main theme of the soundscape was louder (see Table 1). The sound of hammers was a high pitched sound of metal hitting metal, while the sound of power grinders was a constant noise. In the questionnaire, many of the listeners stated that the sound of hammers was the most annoying and disturbing in a generally loud soundscape.

As far as the expressway soundscape is concerned, the lower score can be attributed to the sounds of fast-passing cars which stood out from a generally noisy environment. In this case, there were no sudden and large loudness changes that would distract the listeners, thus the scores were relatively high.

Similarly, when it comes to the stream soundscape, lower results occurred at the moment of the dog's first bark since this sound event was unexpected after a longer period of no loudness changes. After that, the listeners expected similar distracting sounds and, consequently, the scores were higher.

Table 2. The game results for the experimental and control groups

	Children's park		Expressway		Industrial hall		Stream	
Group	EG	CG	EG	CG	EG	CG	EG	CG
Average score	0.86	0.92	0.93	0.96	0.88	0.92	0.94	0.96
Stand. deviation	0.091	0.071	0.058	0.041	0.074	0.056	0.055	0.043
The lowest score	0.47	0.56	0.67	0.79	0.58	0.71	0.83	0.84
N_5 (sone) (7 min)	8.1	6.1	8	6.5	10.5	8.1	9.5	6

6. Future Work

Taking into account the questionnaire and the game results, we can conclude that higher loudness in a certain soundscape is the most annoying factor. It determines general annoyance perception and influences the distraction features of a soundscape considerably. On the other hand, loudness is not the only distraction feature and other sound characteristics must be considered. After being exposed to sudden loudness changes, the listeners adapted, and the only remaining factors of distraction were the sounds perceived as different, unexpected or not corresponding to general soundscape characteristics.

Finally, we can conclude that the results of the testing showed that there is a sophisticated distinction between annoyance and distraction, and that these two terms should be distinguished from each other in the future. Further research on this topic should focus on determining in what way particular sounds differ from their background, as well as identifying the factors that influence human perception which defines these sounds as unexpected for a given soundscape.

One more way that we could proceed is preserving and/or creating ”silent places” within a certain urban soundscape. This research also showed that people prefer the sound of water. This can be a valuable information in soundscape studies in terms of masking the annoying soundscape with the soundscape that consist of water and natural sounds. We could also work on developing innovative design and materials: green spaces, green walls, water walls and other unrecognized ecosystem services.

7. Conclusion

In spite of adapting to a certain sound environment, providing it is not too loud, people still get distracted by unexpected sounds not ”belonging” to the main soundscape print. A soundscape may be loud but, at the same time, not perceived as such; whereas a somewhat different sound in that soundscape, no matter how short, could be the source of annoyance and the reason for distraction.

Finally, we can conclude that urban planning should include zoning requirements for new buildings and constructions, which would offer us the possibility for designing soundscapes.

Acoustic ecology is not just an interesting new aspect of urban studies. In our opinion soundscape designing will be one of the most important part for future city planning in a way to create sustainable and pleasant cities. Natural and water sounds in soundscape can be used as a tool when masking the undesirable sounds and soundscapes.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been supported by the European Union from European regional development fund (ERDF) under the project number KK.01.2.1.01.0103 Acoustical Camera (in Croatian: Akustička kamera).

References

[1] R. M. Schafer, Our Sonic Environment and the Soundscape: The Tuning of the World, Destiny Books, 1994.

[2] H. Fastl, S. Kerber, N. Guzsvány, ”Aspects of startling noises” in Proceedings of Euronoise 2009, Edinburgh, 2009. ISBN: 978-1-61567-680-4

[3] J. Vos, ”The Loudness of Impulse and Road-Traffic Sounds in the Presence of Masking Background Noise”, Acustica – Acta Acustica, **84**(6), 1119-1130, 1998. ISSN 1610-1928

[4] B. Schulte-Fortkamp, A. Fiebig, ”Soundscape Analysis in a Residential Area: An Evaluation of Noise and People’s Mind”, Acta Acustica united with Acustica, **92**(6), 875-880, 2006. ISSN 1610-1928

[5] B. Berglund and M. E. Nilsson, ”On a tool for measuring soundscape quality in urban residential areas”, Acta Acustica united with Acustica, **92**(6), 938-944, 2006. ISSN 1610-1928

[6] J. Y. Jeon. P. J. Lee, J. You, J. Kang, ”Perceptual assessment of quality of urban soundscapes with combined noise sources and water sounds”, Journal of the Acoustical Society of America, **127**, 1357-1366, 2010. https://doi.org/10.1121/1.3298437

[7] A. Torija, D. Ruiz, A. F. Ramos-Ridao, ”Application of a Methodology for Categorizing and Differentiating Urban Soundscapes Using Acoustical Descriptors and Semantic-Differential Attributes”, Journal of the Acoustical Society of America, **134**, 791-802, 2012. https://doi.org/10.1121/1.4807804

[8] A. Ozcevik, Z. Yuksel Can, ”A Laboratory study on Evaluation of Soundscape”, in Proceedings of the Acoustics 2012, Nantes, France, 2012. https://hal.archives-ouvertes.fr/hal-00810900

[9] M. Suhanek, I. Djurek, ”Perception and Evaluation of Sudden Loudness Changes in Urban Soundscapes”, in Proceedings of the 5th Congress of the Alps Adria Acoustics Association 2012, Petrčane (Zadar), Croatia, 1-7, 2012. ISBN 978-953-95097-1-0

[10] E. E. Ryherd, L. M. Wang, ”The Effects of Noise from Building Mechanical Systems with Tonal Components on Human Performance and Perception”, ASHRAE Transactions, AB-10-018 (RP-1322), 2010. http://digitalcommons.unl.edu/archengfacpub/6

[11] L. M. Wang, ”Effects of Building Mechanical System Noise on Worker Performance and Perception”, in Proceedings of NOISE-CON 2010, 2010. http://digitalcommons.unl.edu/archengfacpub/6/

[12] M. Suhanek, I. Đurek, ”Implementation of Bipolar Adjectives pairs in Analysis of Urban Acoustic Environments”, Promet - Traffic & Transportation. **28** (5), 461-470, 2016. https://doi.org/10.7307/ptt.v28i5.2089

[13] E. Zwicker, ”Procedure for calculating loudness of temporally variable sounds”, Journal of the Acoustical Society of America, **62**, 675-682, 1977. https://doi.org/10.1121/1.381580

[14] E. Zwicker, H. Fastl, Psychoacoustics: Facts and models, 2nd ed. Springer-Verlag, Berlin, Heidelberg, New York, p. 220, 1999.
[15] U. Widmann, "A psychoacoustics annoyance concept for application in sound quality", The Journal of the Acoustical Society of America **101**, 3078 (1997); https://doi.org/10.1121/1.418779
[16] S. Viollon and C. Lavandier, "Multidimensional Assessment of the Acoustic Quality of Urban Environments", in Proceedings of Internoise, Nice, France, 2000. ISBN 2951561962, 9782951561960
[17] C. Semidor, "Listening to a City With the Soundwalk Method", Acta Acustica united with Acustica, **92**(6), 959-964, 2006. ISSn 1610-1928
[18] http://www.m-audio.com/, 20.09.2013.
[19] G. Milas, Istraživačke metode u psihologiji i drugim društvenim znanostima (Research Methods in Psychology and other Social Sciences), Naklada Slap, p. 91-118, 153-216, 2009. ISBN 978-953-191-283-9
[20] M. Mejovšek, Metode znanstvenog istraživanja u društvenim i humanističkim znanostima (Methods of Scientific Research in Social Sciences and Humanities), Naklada Slap, p. 71-158, 2007. ISBN 953-191-342-3

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 216-219 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering 2018

ASTESJ

ISSN: 2415-6698

Analysis of Drain Current Transient Response of Gate Pulse Voltage in AlGaN / GaN High Electron Mobility Transistors

Hirohisa Taguchi*, Kazuto Akahori, Takuma Shimazu, Honoka Tanabe

Department of Electrical and Electronic Engineering, School of Engineering, Chukyo University, Nagoya, 4668666, Japan

A R T I C L E I N F O

Article history:
Received: 30 July, 2018
Accepted: 30 September, 2018
Online: 05 October, 2018

Keywords:
AlGaN/GaN HEMTs
Current collapse
Transient response

A B S T R A C T

Recently, several studies focus on a GaN material system that exhibits a significant probability of use in power devices including wide-gap semiconductors. However, the GaN-HEMT is also a structure that easily leads to crystal defects in AlGaN and i-GaN heterojunction. The aim of the study involved investigating the cause of the current collapse in GaN-HEMT after device construction. A GaN-HEMT with a field plate structure was subject to an environmental temperature change from 300 K to 400 K. A pulse voltage was applied to the gate electrode, and the transient response characteristic of the drain current was analyzed. Given the application of the pulse voltage on the gate electrode, charging and discharging of 2 DEG carriers was repeated with respect to crystal defects near the gate electrode. The charge / discharge reduction was observed via a sampling oscilloscope as a transient response. The transient response exhibited an evident dependence on temperature change. The dependence indicated a time constant change, and thus it was possible to calculate the activation energy of crystal defects trapping carriers. The results suggested that the crystal defect evaluation of GaN-HEMT was possible via transient response analysis of 2DEG carrier by using the proposed method.

1. Introduction

The study is an extension of a previous study originally presented in TENCON 2017 [1]. Recently, several studies focus on GaN material system that exhibits a significant probability of use in power devices with respect to wide-gap semiconductors [2]. The GaN exhibits high breakdown voltage characteristics derived from the wide gap structure, and it is possible to exhibit an output exceeding 1 A in the HEMT structure [3]. Extant studies investigated the application of GaN-HEMTs to green ICT systems as high efficiency devices [4]. The device structure of GaN-HEMT is constructed with a superlattice structure of AlGaN and i-GaN. The superlattice structure exhibits a piezo effect and spontaneous polarization effect and generates two-dimensional electron gas at the interface between AlGaN and i-GaN [5]. Thus, a transistor layer that exhibits a low ON resistance is obtained by constructing a channel layer in GaN that should originally correspond to a high resistance material. Therefore, the GaN-HEMT is a power device exhibiting high withstand voltage, high power operation, and low ON resistance. It is expected that it can be potentially utilized in a wide range of applications including power amplification mechanism of mobile phone base station, power units of artificial satellites, and sensor applications in server power supply [6].

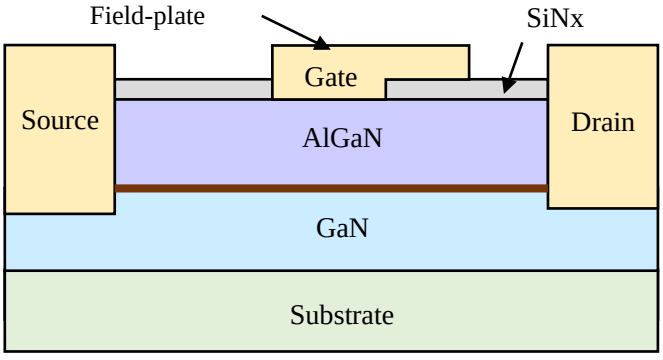


Figure 1 Cross section of the GaN HEMT used in this study

However, the GaN-HEMT is also a structure that easily causes crystal defects in AlGaN and i-GaN heterojunction [7]. A current collapse phenomenon occurs wherein 2DEG carriers are trapped in the crystal defect and the transfer current value sharply decreases with respect to DC response characteristics [8]. Measures including the addition of a surface passivation film

*Hirohisa Taguchi, , Email: htaguchi68@gmail.com

(SiNx), overhanging the direction of the drain electrode of the gate electrode structure (gate plate), and addition of a recess structure between the gate and drain were adopted with respect to the GaN-HEMT structure to suppress the current collapse phenomenon [9, 10]. However, crystal defects are known to increase due to the energization of the device or extreme environmental temperature change, and this deteriorates device characteristics. Additionally, after device construction, it is difficult to identify crystal defects generated in GaN-HEMT [11].

Therefore, the aim of the study involves investigating the cause of current collapse of GaN-HEMT after device construction. A GaN-HEMT with a field plate structure is subject to an environmental temperature change from 300 K to 400 K. A pulse voltage is applied to the gate electrode, and the transient response characteristic of the drain current was analyzed. Given the pulse voltage applied to the gate electrode, charging and discharging of 2 DEG carriers are repeated with respect to crystal defects near the gate electrode. The charge / discharge reduction is observed with a sampling oscilloscope as a transient response. The transient response exhibits an evident dependence on the temperature change. The dependence indicates a time constant change, and thus it is possible to calculate the activation energy of crystal defects trapping carriers. The results suggest that the crystal defect evaluation of GaN-HEMT is possible via the transient response analysis of a 2DEG carrier by using the proposed method.

2. Experimental Method

2.1. Target Device

Figure 1 shows the cross section of the GaN HEMT used in the study. The GaN HEMT includes an i-GaN layer built on a SiC substrate as a channel layer. As a characteristic structure of the HEMT, it includes a field plate that extends the drain side of the T-shaped gate. This suppresses the current collapse phenomenon during normal temperature and high temperature operation.

2.2. Experimental Set Up

Figure 2 shows the experimental diagram in the experiment. In the experiment, the gate voltage was pulsed, and the change in the drain current is observed via a sampling oscilloscope. The base voltage of the gate pulse is set to 0 V, and the rectangular wave pulse is set at the peak voltage of 2 V. The period is 50 ms, and the duty ratio is 0.5. The drain voltage is variable at 0 to 7 V, and the response characteristic observation is set at 3 V. Pulse response characteristics are observed via a sampling oscilloscope (HP83480A, Max: 50 GHz). In order to observe environmental temperature changes, AlGaN / GaN HEMT (Devise under the Test: DUT) is introduced into a nitrogen furnace. The environmental temperature change ranges from 296 K to 373 K. All measurement systems are high frequency compatible (50 GHz).

3. Experimental Results

Figure 3 shows the I-V characteristics of the GaN-HEMT used in the experiment. The blue line in the figure shows the result when only DC voltage is applied. The orange line is obtained by adding high frequency to the gate electrode portion. In the case of only DC voltage, the linear region characteristic of the FET is confirmed immediately after the drain voltage is applied. However, when the drain voltage is 0.2 V or more, the current collapse phenomenon appears, and the current value plummets. When a high frequency component is applied to the gate electrode portion, the carrier trapping effect possessed by crystal defects near the gate electrode is suppressed by the field plate effect, and an almost normal I-V characteristic is obtained. The phenomenon was interpreted as the self-heating of GaN-HEMT in previous studies [12].

Figure 4 shows the time axis response of the drain current when a pulse voltage is applied to the gate electrode portion. The time axis response characteristic in nine points in 10K increments from and Figure. 4 (b) shows the enlarged view of the peak portion where temperature dependence is confirmed. As shown in figure 4

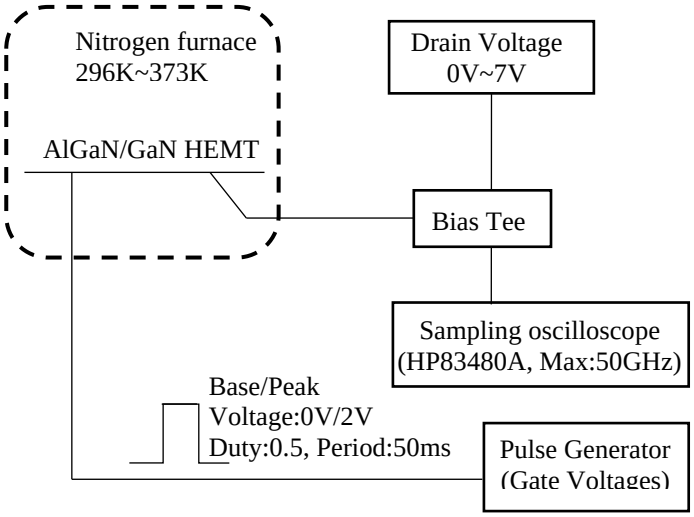


Figure 2 Experimental diagram in this experiment.

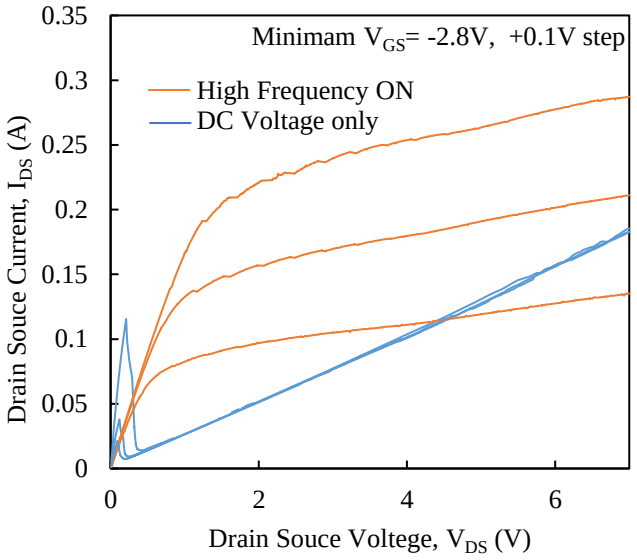


Figure 3: I-V characteristics of the GaN-HEMT.

(a), three types of peak responses are confirmed. Peak 1 in figure 4 (a) in the temperature dependence is not observed. While a gate pulse voltage is applied, a stable 2DEG current exists under the gate, although the 2 DEG current value plummets at the moment

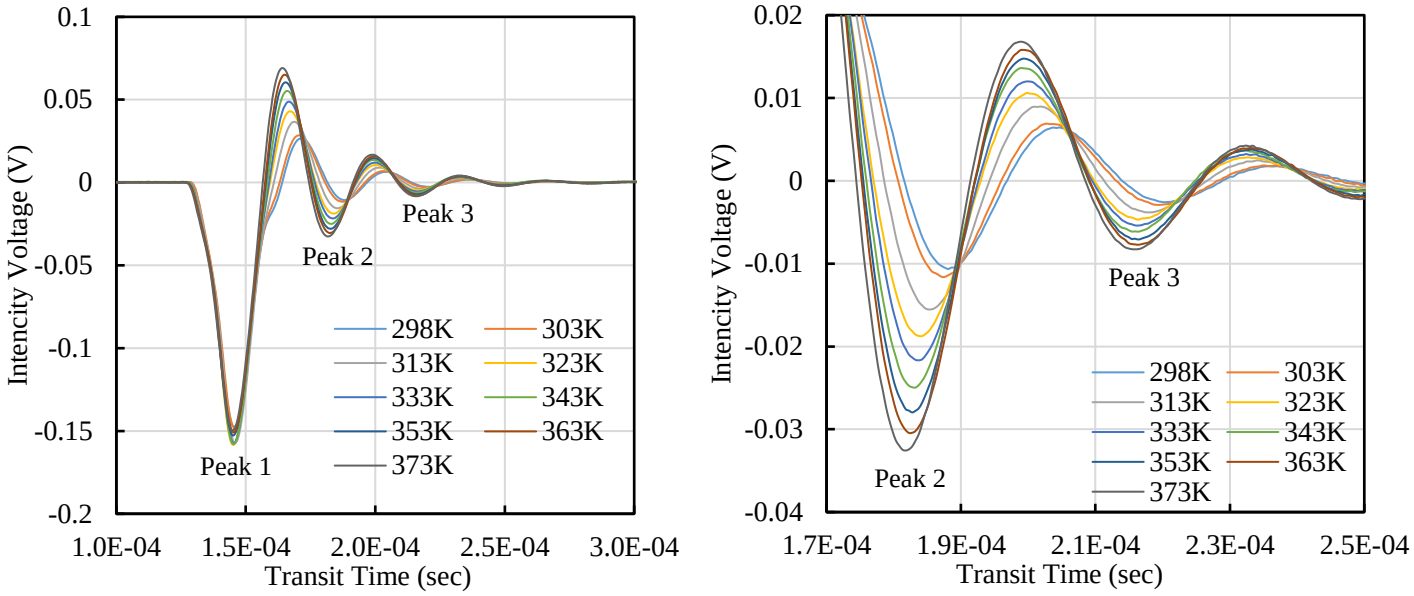


Figure 4 Time-axis response to the drain pulse voltage of the AlGaN/GaN HEMT.
(a) Overall view (b) Enlarged view of peak 2 and peak 3

when the pulse voltage switches to the OFF-State. The peak 1 does not exhibit temperature dependency, and it is considered as a response characteristic when the 2DEG carrier is directly swept from under the gate.296 K is observed by applying a temperature change of up to 373 K. Figure. 4 (a) shows the overall view of the time axis response,

However, peak 2 and peak 3 exhibit an evident temperature dependence as shown in Figure. 4 (b). The temperature dependence in this case exists in terms of both the peak height and peak shape. The peak height indicates the temperature dependence of the number of carriers based on the peak. The peak shape indicates the carrier sweeping process. When the horizontal axis is considered on the time axis, it is interpreted as a transient response. Thus, the shape change indicates a time constant change that limits the transient response. This is discussed in the next section.

4. Discussions

Peak 2 exhibited an evident temperature dependence. The peak height of Peak 2 decreased with increases in the temperature. The peak height was the response voltage and is dependent on the number of carriers. It was considered that the increases in temperature strengthen the piezo effect and increase the thermal carrier supplied from the semiconductor hetero interface. Both the 2 DEG carrier supplied by the voltage and the carrier supplied by the heat were trapped in crystal defects in the AlGaN. The trapping probability was based on the number of carriers wherein thermal carriers increase with increases in the temperature and the number of carriers accumulated in crystal defects also increases [13]. It was considered that the captured carrier is emitted from crystal defects by turning the gate pulse voltage as OFF-State. The emission phenomenon was interpreted as a peak response shown in Figure 4, and transient response characteristics were analyzed.

The response time constant was calculated from the transient response characteristic. As shown in Figure 5, the response time constant decreased with increases in the temperature. Based on the temperature dependence, an Arrhenius plot was executed, and the activation energy ΔE of transient response characteristics was calculated from the slope of the plot. Additionally, ΔE was calculated as approximately 0.575 eV. The value was extremely close to the activation energy of crystal defects in reference [13]. Thus, the transient response indicated by the peak 2 was interpreted as a carrier discharge phenomenon trapped in the crystal defect of the AlGaN layer under the gate.

A pulse voltage was applied to the gate electrode, and thus it was possible to repeat the process of generating / extinguishing the 2 DEG directly under the gate electrode. When only 2DEG carrier generating / extinguished (as shown in Figure 4), peak 2 and peak 3 based on the temperature were not confirmed. When the pulse voltage corresponded to the ON-State at the gate electrode, the 2 DEG carrier was formed under the gate. At the same time, the 2DEG carrier was trapped by crystal defects present in the AlGaN layer near the gate electrode. When the gate pulse voltage turned OFF-State, the 2 DEG carriers were first swept, and the peak 1 was confirmed. There was no reason to observe the temperature dependence in this case. Subsequently, the carriers trapped in crystal defects were gradually released, and peak 2 and peak 3 appeared as transient responses.

We attempted the same analysis for peak 3 although the number of trap carriers was not sufficiently acquired, and the peak intensity was low. An evident temperature dependence was confirmed as shown in the shape in Figure 4. However, an effective analysis result was not obtained. Evidently, the response time of peak 3 was slower than that of peak 2. Additionally, the peak height was lower than that of peak 2. Given that the carrier trap to the crystal defect formed peak 3, a period of time elapsed

in releasing the carrier, and it was considered as the characteristic of the crystal defect in the deep part of the crystal. The possibility was considered wherein the presence of crystal defects in the i-GaN crystal depth builds a channel layer [14].

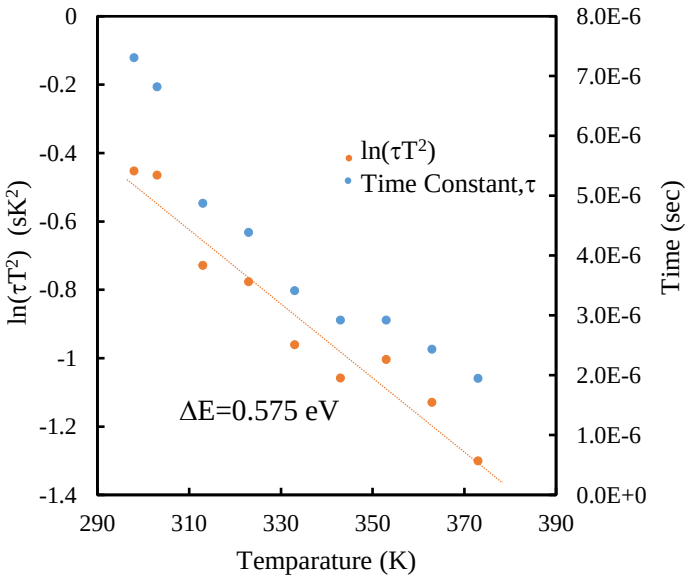


Figure 5 Arrhenius plot using the time constant.

In the study, a pulse voltage was applied to the gate electrode of the completed GaN-HEMT, and the transient response change of the drain current was precisely observed and analyzed. This was a simple method. However, it enabled the evaluation of crystal defects in the completed GaN-HEMT. Although the ICTS method and the DLTS method were used for the same analysis, the method was an extremely simple method when compared with the aforementioned methods and exhibited excellent versatility.

5. Conclusions

The aim of the study involved investigating the cause of current collapse of GaN-HEMT after device construction. A change in the environmental temperature from 296 K to 373 K was applied to GaN-HEMT with a field plate structure. A pulse voltage was applied to the gate electrode, and the transient response characteristic of the drain current was analyzed. Given the application of the pulse voltage to the gate electrode, charging and discharging of 2 DEG carriers was repeated with respect to crystal defects near the gate electrode. The charge / discharge reduction was observed with a sampling oscilloscope as a transient response. The transient response exhibited an evident dependence on temperature change. The response time constant is calculated from the transient response characteristic. As shown in Figure 5, the response time constant decreased with increases in the temperature. Based on the temperature dependence, an Arrhenius plot was executed, and the activation energy ΔE of transient response characteristics was calculated from the slope of the plot. Additionally, ΔE was calculated as approximately 0.575 eV. The value was extremely close to the activation energy of crystal defects in extant studies. The dependence indicated a time constant change, thereby facilitating the calculation of the activation energy of crystal defects trapping carriers. It was suggested that the crystal defect evaluation of GaN-HEMT is possible via transient response analysis of 2DEG carrier by the method.

Acknowledgment

The authors are grateful to the Chukyo University Research Foundation for financial assistance.

References

[1] A. Otake, S. Nakano, K. Yamanaka, H. Taguchi, “Temperature dependence of drain current transient response by gate pulse voltage in AlGaN / GaN high electron mobility transistors” TENCON 2017 IEEE Region 10 Conference, Penang, Malaysia, 2017. https://doi.org/ 10.1109/TENCON.2017.8228128

[2] S. Nakamura, T. Mukai, M. Senoh, “Candela-class high-brightness InGaN/AlGaN double-heterostructure blue light-emitting diodes” Appl. Phys. Lett. 64(13), 1687–1689, 1994. https://doi.org/10.1063/1.111832

[3] G. Nishio, K. Nakatani, T. Ishizaki, “Design of harmonic processing circuit for microwave GaN-HEMT power amplifier”, IEEE International Meeting for Future of Electron Devices, Kansai (IMFEDK), Kyoto, Japan, 2015. https://doi.org/10.1109/IMFEDK.2015.7158583

[4] R. Quay, D. Schwantuschke, E. Ture, F. van Raay, C. Friesicke, S. Krause, S. Müller, S. Breuer, B. Godejohann, P. Brückner, “High-power microwave GaN/AlGaN HEMTs and MMICs on SiC and silicon substrates for modern radio communication” Phys. Status Solidi A, 215(9), 1700655(1–7), 2018. https://doi.org/10.1002/pssa.201700655

[5] K. Shinohara, C. King, A. Carter, E. Regan, A. Arias, J. Bergman, M. Urteaga, B. Brar “GaN-based field-effect transistors with laterally gated two-dimensional electron gas” IEEE Electron Device Letters, 39(3), 417–420, 2018. https://doi.org/10.1109/LED.2018.2797940.

[6] H. Watanabe, T. Fukami, H. Saito, A. Tomiki, O. Ceylan, H. Nunomura, O. Shigeta, T. Shinke, K. Kojima, “High speed downlink system for small satellite and high-efficiency x-band GaN SSPA” IEEE MTT-S International Microwave Symposium (IMS2014), Tampa, FL, USA, 2014. https://doi.org/10.1109/ MWSYM.2014.6848670

[7] H. Sasaki, T. Hisaka, K. Kadoiwa, T. Oku, S. Onoda, T. Ohshima, E. Taguchi, H. Yasuda, “Ultra-high voltage electron microscopy investigation of irradiation induced displacement defects on AlGaN/GaN HEMTs” Microelectronics Reliability, 81, 312–319, 2018. https://doi.org/10.1016/j.microrel.2017.10.005.

[8] H. Huang, Y. Liang, G. Samudra, T. Chang, C. Huang, “Effects of gate field plates on the surface state related current collapse in AlGaN/GaN HEMTs” IEEE Transactions on Power Electronics, 29(5), 2164–2173, 2013. https://doi.org/ 10.1109/TPEL.2013.2288644.

[9] W. Waller, M. Gajda, S. Pandey, J. Donkers, D. Calton, J. Croon, S. Karboyan, J. Šonský, M. Uren, M. Kuball, "Impact of silicon nitride stoichiometry on the effectiveness of AlGaN/GaN HEMT field plates" IEEE Transactions on Electron Devices, 64(3), 2017. https://doi.org/10.1109/TED. 2017.2654800

[10] D. Čučak, M. Vasić, O. García, J. Oliver, P. Alou, J. Cobos, A. Wang, S. Martín-Horcajo, M. Romero, F. Calle, “Physics-based analytical model for input, output, and reverse capacitance of a GaN HEMT with the field-plate structure” IEEE Transactions on Power Electronics, 32(3), 2017. https://doi.org/10.1109/TPEL.2016.2569404

[11] S. Huang, Q. Jiang, S. Yang, C. Zhou, K. Chen, “Effective passivation of AlGaN/GaN HEMTs by ALD-grown AlN thin film”, IEEE Electron Device Letters, 33(4), 516–518, 2012. https://doi.org/10.1109/LED.2012.2185921.

[12] C. Wang, Y. Xu, X. Yu, C. Ren, Z. Wang, H. Lu, T. Chen, B. Zhang, R. Xu, "An electrothermal model for empirical large- signal modeling of AlGaN/GaN HEMTs including self-heating and ambient temperature effects" IEEE Transactions on Microwave Theory and Techniques, 62(12), 2014. https://doi.org/10.1109/TMTT.2014.2364821.

[13] M. Mielenz, J. Brox, S. Kahra, G. Leschhorn, M. Albert, T. Schaetz, H. Landa, B. Reznik, “Trapping of topological-structural defects in Coulomb crystals” Phys. Rev. Lett. 110, 2013. https://doi.org/10.1103/ PhysRevLett.110.133004.

[14] J. Joh, J. Alamo, “A current-transient methodology for trap analysis for GaN high electron mobility transistors”, IEEE Trans. On Electron Devices, 58(1), 132–140, 2011. https://doi.org/10.1109/TED.2010.2087339.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 220-225 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Development and Testing of Intelligent Wheelchair Controller for Quadriplegic Patients

Mohammed Faeik Ruzaij Al-Okby [*,1], Sebastian Neubert [2], Norbert Stoll [2], Kerstin Thurow [3]

[1]*Technical Institute of Babylon, Al-Furat Al-Awsat Technical University (ATU), Babylon, Iraq*

[2]*Institute of Automation, University of Rostock, Rostock 18119, Germany*

[3]*Center for Life Science Automation (celisca), University of Rostock, Rostock 18119, Germany*

ARTICLE INFO

Article history:
Received: 09 August, 2018
Accepted: 30 September, 2018
Online: 05 October, 2018

Keywords :
Intelligent Wheelchair
Quadriplegia
Orientation Detection

ABSTRACT

In this research paper, the development and evaluating of a smart controller for electrical powered wheelchairs are presented. The controller aimed to assist quadriplegic, paralyzed, and handicap patients who cannot use their hands to drive an electrical wheelchair by using a joystick controller. Two sub control units have been combined in one hybrid system to create the current version which are voice, and head tilt controllers. They are activated and operated in parallel at the same time to allow the wheelchair user to choose the preferred control method. The voice controller is activated by the user's voice command to control the wheelchair instead of a joystick. The head tilt controller uses the user's head motions to create control commands instead of the joystick controller. The head tilt controller design is based on using two embedded MEMS orientation detection modules as input measurement units. The system uses a modern low power consumption microcontroller to analyze the received information and data from inputs and creating the prompt control commands and send it to the wheelchair motors driver as an output unit.

1. Introduction

The use of assistive and rehabilitation applications has increased rapidly in recent years because of the huge increase in the numbers of handicapped and paralyzed people. In the traditional wheelchair, joysticks controller has been used to generate the control commands, which requires good control of the user's hand. However, quadriplegic and paralyzed patients cannot use a conventional joystick controller. Thus, several new technologies, methods, and inventions have to be developed to serve this class of patients and to reduce their sufferance. Nowadays, several input signals from the user's body can be used such as body gesture and motions [1-3], brain electrical signals [4-6], user voice commands [7-9], and electrical activity in the body muscles (EMG) [10-11]. Several approaches have been used for each signal. Each approach has its advantages and disadvantages, more details and information are available in [12-13]

In this paper, the design, and evaluation of the hybrid controller for electrical wheelchairs are explained. The controller design takes into consideration the target users' requirements. The target users lost the ability to use their upper and lower limbs and cannot handle a conventional joystick controller. The novelty of the proposed system is based on using more than one controller flexibly to allow the user choosing the preferred controller. The novel implementation of the voice controller allows the system to avoid false positive (FP) errors by using two different voice recognition modules combined by a false positive errors cancelation algorithm. The head tilt controller has a unique design to perform several tasks based on using many functions to improve the performance of the system in different situations such as applied emergency stop or driving the system in ramps and non-straight roads. The system has been tested with the implementation of several tasks, some important factors such as response time and stop distance have been measured to check the new system controller's performance in comparison to an original joystick controller. The combination of several sensors and detection units in the voice and head tilt controllers produce a durable and simple to use system that helps the target users to enhance the quality of daily life. The proposed work is an improvement and a development of previous work that can be found in [14-15].

2. Methodology

There are three main layers of the system which are the input layer, the main processor, and the execution platforms. The system main layers have been explained in Fig. 1. The input layer

[*]Mohammed Faeik Ruzaij Al-Okby, Technical Institute of Babylon, Al-Furat Al-Awsat Technical University (ATU), Babylon, Iraq. Phone: +9647722365739. Email: Mohammed.al-Okby@celisca.de, abnalfratain@yahoo.com.

includes all the sensors and modules that are responsible for picking up the control signals from the user side and sending it to the main processing unit. The main processor analyses and processes the data and information of the sensors and modules in the input layer. It is responsible for the generation and sending of the control commands to the output layer which is also called the execution platforms. The following section explains each layer in detail:

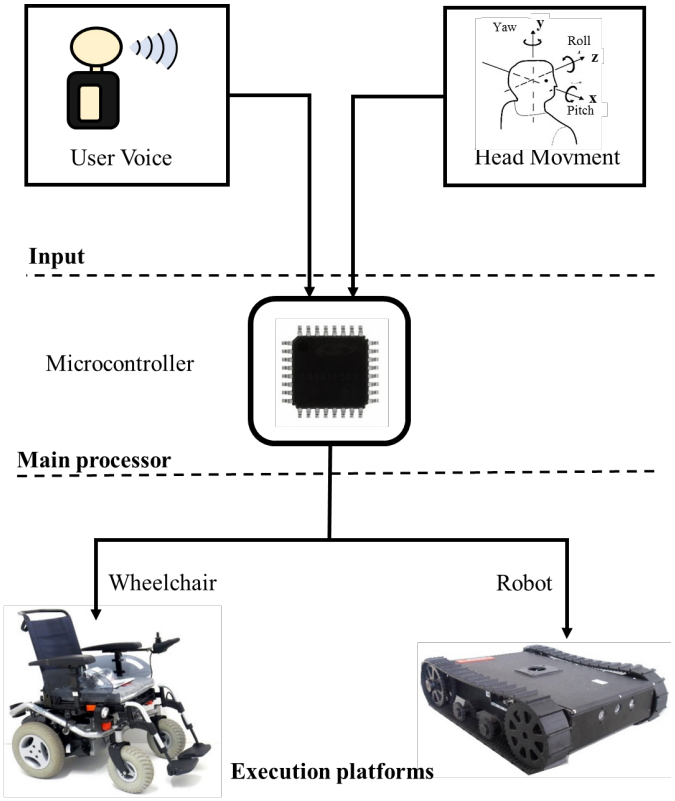


Figure 1: System structure.

2.1. Input Layer

The input layer deals with the signals and parameters picked up from the user body to generate specific control commands. It has two main signals from the wheelchair users which are the user voice command and the user's head motions. There are two units for signal acquisition which are the voice recognition (VR) unit and the orientation detection (OD) unit. These two units are the main core of the voice and head tilt controllers.

The VR unit is a combination of two sophisticated modules using different voice recognition processors and algorithms. This makes the voice controller more reliable and allows the system to detect and eliminate the FP errors. The two voice recognition modules work together in parallel and send the VR results to the microcontroller as command index value using a universal asynchronous receiver/transmitter (UART) bus or by sending a binary value for a specific input/output port pin.

The first voice recognition module used in the voice controller is the Easy VR module which uses the strong voice processor RSC-4128 from Sensory, USA. The Easy VR module has the ability to operate in Speaker Dependent SD as well as Speaker Independent SI modes. The SD means that the system detects the voice command depending on the individual characteristic of the voice command. This makes the system respond only to the person who was trained before to the system. SI mode means, the system responses to any user giving the selected voice commands. For SD mode, the module uses the well-known Dynamic Time Warping (DTW) algorithm. The module uses the hidden Markov model (HMM) for the SI mode. The module can be trained with up to 32 voice commands.

The second module is the SpeakUP Click™ module (MikroElektronika, Serbia). It only uses the DTW algorithm and is operating in SD mode only. This module can work standalone or with the host processor. It can be trained with up to 200 voice commands. More details regarding the used VR module can be found in [16-17].

The second acquisition unit in the input layer consists of two modern orientation detection sensors from Bosch sensor tech. Inc., Germany called BNO055. Both modules are identical and operate at the same bus speed and settings. They use I^2C bus to communicate with the host processor; each module uses its own I^2C address. The two modules are fixed in different locations on the system depending on the required orientation measurement. The first module is used to generate the motion control command. A traditional PC headset has been used to fix the first module in the middle of the headset. The second BNO055 module is used to detect the road slope or wheelchair reference orientation. It is fixed on the wheelchair chassis. The internal structure of BNO055 module includes 4 MEMS sensors combined with an ARM 0 microcontroller. The MEMS sensors are a gyroscope, magnetometer, temperature sensor, and accelerometer. A data fusion algorithm has been used to analyze the data from the four sensors and feeds the host processor by a ready to use orientation information in different measurement styles such as quaternion vector, linear acceleration, rotation vector, and Euler angles. In current work, the angle degree measurements of the Euler angles Pitch, Roll, and Yaw have been selected as input to the processing unit to represent the user's head and reference orientation of the wheelchair [18].

2.2. Main processing Layer

The main processing layer interface between the input and output layers. The received information and data from the input unit are transported to the main processing unit using several communication busses and ports. The received data represent the user commands and is processed by the main microcontroller using a specific algorithm depending on the type of sensor sending the information. The results from the microcontroller will be converted to a specific hex-decimal message to control the wheelchair components. The Silicon labs EFM32GG990F1024 microcontroller has been used as the main controller of the system. Several sufficient input/output ports with wide types of

communication buses have been embedded such as UART, I²C, USB, and general-purpose input-output pins GPIO [19].

One of the important tasks of this layer is to prevent the conflict between the modules and sensors in the input layer. The head tilt controller is activated as main controller when the system starts. The voice controller is activated for complementary functions such as controlling speed level and turning on and off lights and signals. The user can select the voice controller as main motion controller by giving the voice command "voice" and he can return to the head tilt as main controller using the voice command "orientation". Fig. 2 shows the system flowchart and the transition between the sub-controllers.

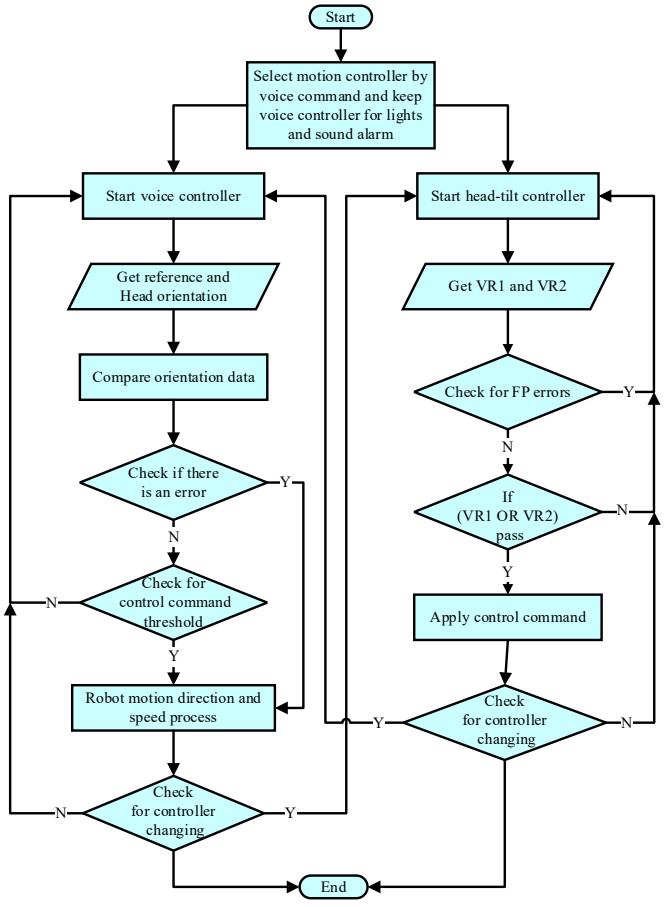


Figure 2: System flowchart

2.3. Output Layer

The output layer is responsible for the execution of the control commands. It is implementing motion commands for controlling the wheelchair electrical motors by changing the motors rotation direction and speed. In the presented system, the electrical powered wheelchair Meyra Smart 9.906 (MEYRA GmbH, Germany) is used as output or execution layer. The robotic motor driver Sabertooth 2x32 Amp (Dimension Engineering LLC., USA, see Fig. no. 3) has been used to control the wheelchair motors instead of original VR2 motor driver unit. The same motor driver unit has been used in jaguar lite robot and H20 robot which enables to use the system in robotic applications, too.



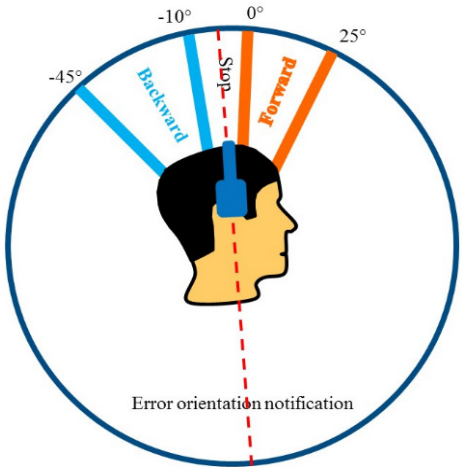
Figure 3: Sabertooth 2x32 Amp motor driver

3. System Description

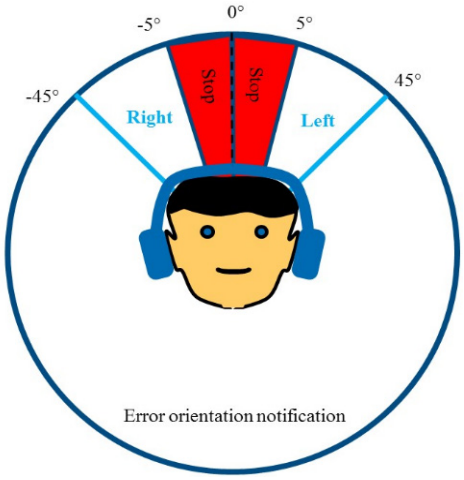
In the presented work, the two sub-controllers are activated together to cover all the required wheelchair tasks. The head tilts controller cannot cover all the control commands of the wheelchair. It is mainly used for motion and speed adjustment control commands and not covers other commands like turning on and off lights or sounding alarms and signals. The voice controller was tested at different noise levels for several voice commands in [9], [14]. In this current work, it is used as an assistive controller with a few commands for controlling additional parameter.

The head tilt controller has two main orientation measurements, the first is the orientation of the user's head which is responsible for generating the motion control commands. The generation of the motion commands depends on the value and the direction (the axis) of the head tilt angle. The second orientation measurement is for the reference wheelchair orientation. This measurement is used to calibrate the system in case of passing non-straight roads. The output of the orientation modules has been set in the form of Euler angles. Three Euler angles have been used to represent the body orientation around the main axes x,y,z using the tilt angles pitch (P), roll (R), and yaw (Y) respectively. The modules send the orientation data via I²C bus to the main microcontroller, which receives and processes the head and reference orientation information. The motion control commands are generated by specific motions of the user's head. The speed and direction of the control command are determined based on the tilt axis for direction and the angle degree value for the speed. There are five main motion commands for controlling the presented system. Each command has specific regions of tilt angles to control the wheelchair motion. The forward command starts and ends by tilting the user head between (0°<P<25°) and (5°<R<45° and -45°<R<-5°). The left and right commands start from 0 to maximum speed by tilting the user head between (-10°>P>0°) and (5°<R<45° and -45°<R<-5°). The stop region is located between (-10°>P>0°) and (-5°>R>5°). The backward command tilt angles are located between (-10°>P>-45°) and (5°<R<45° and -45°<R<-5°). All other tilt angles are not used for generating control commands. This region is used for activating error messages in the system display or for activating emergency stop functions to protect the user and the system in specific cases such as the headset fall down or unexpected collision. Fig. 4

shows the programmed head tilt angles for the roll and pitch Euler angles.



a) Pitch (P)



b) Roll (R)

Figure 4: Control commands regions

4. Experimental Results

4.1. System Indoor Test

Several tasks have been tested to check the performance of the hybrid system. The tasks are divided into an individual task for each controller and for both controllers together. A group of complementary voice command has been tested such as "faster = increase the tilt to speed ratio", "slower = set the tilt to speed to the basic ratio", "light off = turn lights off", "light on = turn lights on", and "stop". The head tilt controller was tested with the motion commands. 14 users tested the system with several tasks. All the 14 users are normal user without disabilities. The users used the system as quadriplegics patient without using their upper and lower limbs. Each task was repeated 10 times by all users. The tested tasks included the most important functions of the system. The tasks have been performed in a known path including a narrow corridor and a furnished laboratory. The test had two parts. In the first part, the wheelchair should be driven in a 1.5 m wide narrow corridor. In this part, the user should give the voice command "faster" to change the head tilt controller sensitivity from the primary speed 0°-25° P tilt angle → 0-2 km/h to the medium speed 0°-25° P tilt angle → 4 km/h (first task T1). In the second task, the user gives the voice command "light on" to activate the system lights T2. When the user reaches the laboratory door, he reduces the wheelchair speed sensitivity to the basic level using the voice command "slower" T3. Inside the laboratory, the user executes another group of tasks. The user needs to pass the laboratory entrance and avoid the collision with some obstacles (see Fig no. 5) T4. After driving the system and passing narrow corridor, the user needs to activate the emergency stop function of the head tilt controller by tilting his head beyond the programmed ranges for pitch or roll Euler angles T5. Then, the user needs to drive the wheelchair to the second laboratory door for exit. Before exiting, the user deactivates the system lights using the voice command "light off" T6. Finally, the user should exit the laboratory and stop the wheelchair in the outer corridor using "stop" voice command T7. The navigation map of the system tests is shown in Fig. 5.

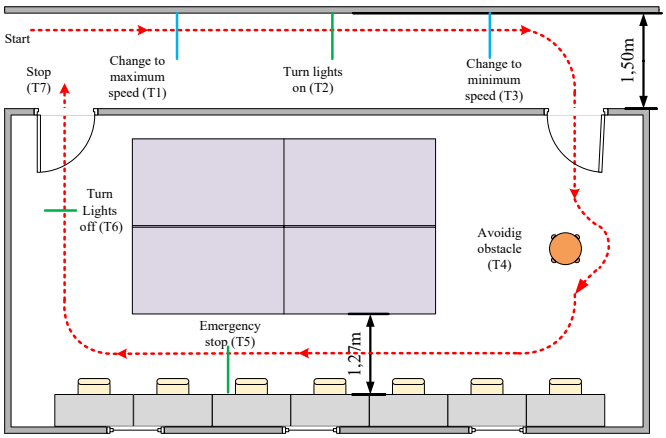


Figure. 5: System tests map

Table 1 includes the system indoor test results in the form of correct command execution (accuracy), the number of false positive error FP, and the number of false negative error FN for every task. The results clearly show that the head tilt controller performance is better than voice controller with 0 % errors and 100 % correct commands accuracy.

Table 1: Indoor test results in the form of correct command execution

Task	Accuracy	FP	FN
T1	95.71 %	0	6
T2	96.42 %	0	5
T3	95.71 %	0	6
T4*	100 %	0	0
T5*	100 %	0	0
T6	97.85 %	0	3
T7	96.42%	0	5

* Head tilt commands

4.2. Reaction time test

The reaction time in this test refers to the time period between giving the command by the user and the complete execution of

the command by the system in real time. Two parameters have been measured in this test which are the time and the distance. The three controllers, head, voice, and joystick were tested in the same environment with identical test parameters and conditions such as user weight, wheelchair speed, and the surrounding noise in the test environment. A fixed ≈ 1.716 km/h wheelchair speed was selected for the test for all three controllers. The wheel rotation per minute (RPM) was calculated and adjusted to 27.6 RPM individually for each controller using the digital tachometer DT-2234C. Only one user with a weight of app. 85 kg performed the test for the three controllers. All tests have been taken at the same laboratory with a surrounding noise of ≈ 54 dB.

The reaction time and the distance to stop the system are calculated using video frames from a high-speed camera. The camera can record the test video in a range from 30 to 1000 frames per second. The measured time period starts when the user gives the stop control command and finishes when the wheelchair stopped completely. The test has been repeated 10 times for each controller and the average reaction times and distances have been calculated. The same steps repeated identically for the head tilt, joystick, and voice controllers. The "stop" command was given in different ways depending on the used controller. The joystick controller has a specific joystick range for stop command which can be easily recorded by the camera. The voice controller will stop the system when the user gives the voice command "stop" and to make it easy to be recorded by the camera the user was asked to lift his hand up when he gives the "stop" command. Finally, for the head tilt controller, the camera can easily record the stop command which is activated by returning the user's head to the range (-10°>P>0°, -5°<R<5°). Fig. 6 illustrates the capturing of the stop command frame for all controllers.



a) Head tilt controller



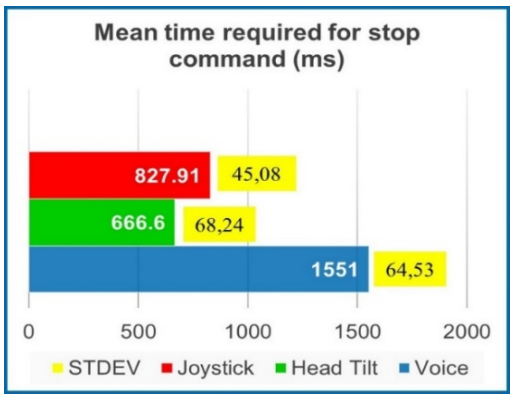
b) Joystick controller



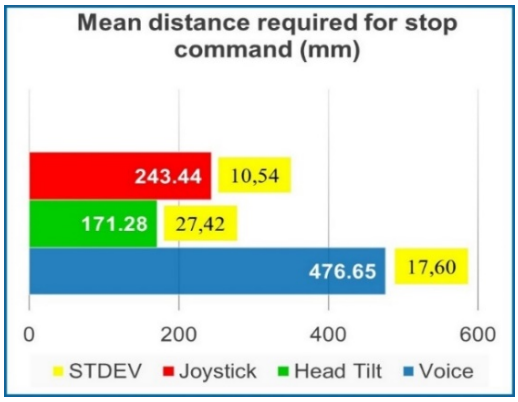
c) Voice controller

Figure 6: Stop command frames of the different controllers

Fig. 7 explains the statistical chart of the calculated reaction time and stop distance of the three controllers. The charts show the mean values for the measured parameter with the standard deviation (STDEV) for 10 samples for each controller. The results showed that the head tilts controller performance is better compared to the other controllers for the two calculated parameters.



a) Reaction time



b) Stop distance

Figure. 7: Reaction time and distance for the stop command

4.3. Questionnaire

For the evaluation of the system usability, the fourteen users in the previous tests were asked seven questions about their feeling, adaptation, and opinions about the system. Each question should get a numerical evaluation from 0 minima up to 10 maximums. Table 1 summarizes the evaluation questions. Fig. 8 and Table 2 explain the questionnaire results.

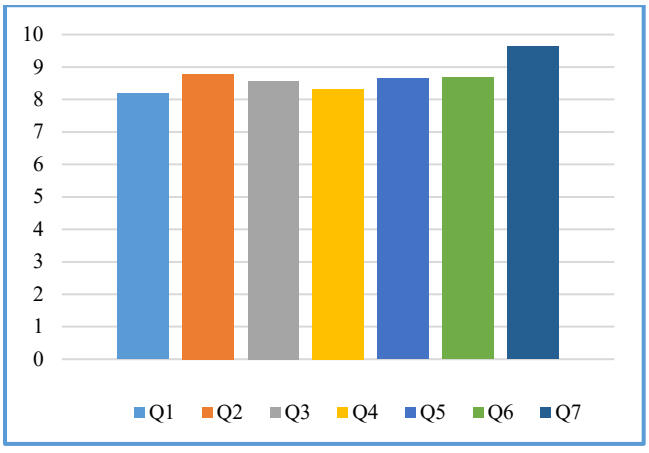


Figure 8: Questionnaire result for the system

Table 2: Questionnaire evaluation questions

Question	Question
Q1	Is it easy to use head tilt controller?
Q2	How do you find the comfortability of wearing a headset?
Q3	How do you find the reaction time for head tilt controller?
Q4	How do you find the reaction time for voice controller?
Q5	How do you find the voice command response?
Q6	Is it easy to pass narrow area?
Q7	Is it easy to control lights and signals?

Table 3: Questionnaire results

User	Q1	Q2	Q3	Q4	Q5	Q6	Q7
U1	7,5	8	7,5	7,5	9	10	10
U2	7,5	8	7	9	7,5	10	10
U3	9	10	10	8	10	7,5	10
U4	8,5	9	8	7,5	9	9	10
U5	7	8	7,5	7	7,5	8	9
U6	7,5	9	10	9	10	8	10
U7	9	8	10	10	8	10	10
U8	9	8	9	10	8	8	10
U9	8	9	10	9	10	8	10
U10	9	10	10	8	10	8	10
U11	8,5	9	8	7,5	9	9	10
U12	8	10	8	8	8	9	9
U13	9	9	8	8,5	7	10	8
U14	7,5	8	7	8	8	7,5	9
Mean	8,214	8,785	8,571	8,3573	8,642	8,714	9,642
SDV	0,726	0,801	1,206	0,928	1,063	0,974	0,633

The abbreviation STDEV refers to the standard deviation. The results of the questionnaire show differences in the system evaluation between the users. The highest mean value with the minimum standard deviation was observed for question Q7 ("Is it easy to control lights and signals?"). The highest difference between the users' answers with a standard deviation of 1.206 was observed for Q3 ("How do you find the reaction time for head tilts controller?"). This may result from the fact, that the users were previously informed about the command confirmation function of the head tilts controller that makes a 100 ms delay in the starting of each motion command. The abbreviation STDEV refers to the standard deviation.

5. Conclusions

In this paper, the development and testing of a hybrid wheelchair controller have been described. The system consists of two sub-controllers which are the voice and head tilts controller. The voice controller uses two voice recognition modules combined to enhance the recognition accuracy and reduce the errors. The head tilts controller tests revealed a good performance in indoor as well as outdoor tests. The tests show that the head tilts controller is easier and more accurate in wheelchair controlling than the voice controller and it is thus selected to be the main controller. The combination of voice and head tilt controllers make the system more flexible and covers all wheelchair functions such as turning lights, sound alarms, and signals on and off. It also enables the selection of the speed range.

Acknowledgment

The presented work has been supported by the Iraqi Ministry of Higher Education and Scientific Research and the German Academic Exchange Service (DAAD, Germany).

References

[1] Henrik Vie Christensen A and Juan Carlos Garcia B,"Infrared Non-Contact Head Sensor, for Control of Wheelchair Movements," Book Title Book Editors IOS Press, 2003.

[2] Farid Abedan Kondori, Shahrouz Yousefi, Li Liu, Haibo Li, "Head Operated Electric Wheelchair", in Proc. 2014 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), San Diego, CA, USA, 6-8 April 2014, pp. 53-56.

[3] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Auto Calibrated Head Orientation Controller for Robotic-Wheelchair Using MEMS Sensors and Embedded Technologies", in Proc. 2016 IEEE Sensors Applications Symposium (SAS 2016), Catania, Italy, 20-22 April, 2016, pp. 433-438.

[4] Kazuo Tanaka, Kazuyuki Matsunaga, and Hua O. Wang, "Electroencephalogram-Based Control of an Electric Wheelchair", IEEE Transactions on Robotics, Vol. 21, No. 4, August 2005.

[5] K. Tanaka, K. Matsunaga, and H. O. Wang, "Electroencephalogram-Based Control of an Electric Wheelchair," IEEE Transactions on Robotics, Vol. 21, No. 4, pp. 762–766, 2005.

[6] I. Iturrate, J. Antelis, and J. Minguez, "Synchronous EEG brain-actuated wheelchair with automated navigation," in Proc. of IEEE International Conference on Robotics and Automation (ICRA '09), Kobe, Japan, 2009, pp. 2318–2325.

[7] Masato Nishimori, Takeshi Saitoh, and Ryosuke Konishi, "Voice Controlled Intelligent Wheelchair", in Proc. SICE Annual Conference 2007, Sept. 17-20, 2007, Kagawa University, Japan, pp. 336-340.

[8] Aruna. C, Dhivya Parameswari. A, Malini. M and Gopu. G, "Voice Recognition and Touch Screen Control Based Wheelchair for Paraplegic Persons", in Proc. Green Computing, Communication, and Electrical Engineering, Coimbatore, 2014, pp. 1-5.

[9] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Hybrid Voice Controller for Intelligent Wheelchair and Rehabilitation Robot Using Voice Recognition and Embedded Technologies", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 20 No. 4-2016, Fuji Technology Press Ltd, Tokyo, Japan, pp. 615-622.

[10] Satoshi Ohishi and Toshiyuki Kondo, "A Proposal of EMG-based Wheelchair for Preventing Disuse of Lower Motor Function", in Proc. Annual Conference of Society of Instrument and Control Engineers (SICE), August 20-23, 2012, Akita University, Akita, Japan, pp. 236-239.

[11] Z. Yi, D. Lingling, L. Yuan, and H. Huosheng, "Design of a surface EMG based human-machine interface for an intelligent wheelchair," in Proc. of 2011 10th International Conference on Electronic Measurement Instruments (ICEMI), Chengdu, China, 2011, pp. 132–136.

[12] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow,"Design and Testing of Low Cost Three-Modes of Operation Voice Controller for Wheelchairs and Rehabilitation Robotics", in Proc. 9th IEEE International Symposium on Intelligent Signal Processing WISP2015, Siena, Italy, May 15 -17, 2015, pp. 114-119.

[13] Mohammed Faeik Ruzaij Al-Okby, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Low-Cost Hybrid Wheelchair Controller for Quadriplegias and Paralysis Patient", Advances in Science, Technology and Engineering Systems Journal (ASTESJ), Vol. 2, No. 3, pp. 687-694, 2017.

[14] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Multi-Sensor Robotic-Wheelchair Controller for Handicap and Quadriplegia Patients Using Embedded Technologies", ", in Proc. 2016-9th International Conference on Human System Interactions (HSI), Portsmouth, United Kingdom, 6-8 July 2016. pp.103-109.

[15] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Design and implementation of low-cost intelligent wheelchair controller for quadriplegias and paralysis patient", in Proc. 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 26-28 January 2017, pp. 399-404

[16] Speak Up Click User Manual Ver.101, MikroElektronika, Belgrade, Serbia, 2014.

[17] EasyVR 2.0 User Manual R.3.6.6., TIGAL KG, Vienna, Austria, 2014.

[18] BNO055 data sheet, [online]. Available: https://ae-bst.resource.bosch.com/media/_tech/media/datasheets/BST_BNO055_DS000_14.pdf. [Accessed: 15 Jun 2016].

[19] EFM32GG990 DATASHEET, [Online]. Available: http://www.silabs.com. [Accessed: 20 December 2015].

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 224-229 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Dual-Band 90-Degree SiGe HBT Active Phase Shifter Based on Band-Pass and Band-Stop Designs Using Dual-Band Resonators

Yasushi Itoh*

Shonan Institute of Technology, Electrical and Information Engineering, 251-8511, Japan

ARTICLE INFO

Article history:
Received: 22 June, 2018
Accepted: 14 July, 2018
Online: 05 August, 2018

Keywords:
Microwaves
Phase Shifter
Differential Amplifier
Dual-Band

ABSTRACT

A dual-band 90-degree SiGe HBT active phase shifter based on band-pass and band-stop designs using dual-band resonators is presented in this paper. The active phase shifter employs differential configuration and has band-pass and band-stop filters in the load circuit. By switching two output ports of the differential amplifier, 90-degree phase shifting has been realized at dual bands. The implemented dual-band active phase shifter using 0.35 micron SiGe HBT has achieved a gain of 9.3 dB and a phase shift of 95 degrees at 0.74 GHz as well as a gain of 8.9 dB and a phase shift of 98 degrees at 0.88 GHz. This is the first paper to present a dual-band active phase shifter based on band-pass and band-stop designs using dual-band resonators.

1. Introduction

This paper is an extension of the work originally presented in the 47th European Microwave Conference, 2017 [1]. Recently, multi-band phase shifters are being developed for the next generation multi-band and multi-mode wireless communication systems using active phased array modules and/or beam-forming networks [2]-[3]. Various types of multi-band phase shifters have been reported, including reflection-type [4]-[5], loaded-line type [6]-[7], vector-sum type [8] and wideband distributed amplifiers with varactor-tuned LC networks [9]-[10]. However, there was no report on the low-pass high-pass phase shifters for the multi-band or multi-mode applications. In order to address this issue, a novel dual-band active phase shifter using band-pass and band-stop designs has been presented in [1]. The active phase shifter is based on a differential amplifier and employs two types of different loads, that is, band-pass or band-stop filters. With the use of abrupt phase shifting characteristics at around cutoff frequencies of the filters [11]-[12], variable phase shifting can be easily achieved at multiple bands by switching band-pass and band-stop filters. Although the conventional low-pass high-pass phase shifters [13]-[14] are limited to single-band operation, band-pass and band-stop designs can make multi-band operation available. In [1], 90-degree phase shifting has been successfully achieved at dual-band by switching two output ports of the differential amplifier. From the viewpoint of multi-band phase shifting as well as high gain, the active phase shifter is considered to be one candidate for achieving various phase shifting at multiple bands, which would be used in the next generation multi-band or multi-mode wireless communication systems with active phased array antenna and beam forming networks. To extend the work [1], the dual-band passive phase shifter based on band-pass and band-stop designs is newly described, showing that the dual-band 90-degree phase shifting is unavailable because of a low Q-factor of the circuit elements. Then how to address this design issue is presented by utilizing 180-degree out of phase performance of the differential amplifier. It is also added in description that an impedance matching has to be taken into account since the band-pass and band-stop filters using dual-band resonators are actually employed in the output matching circuit of the differential amplifier. The outstanding features of the multi-band active phase shifter based on band-pass and band-stop designs over the previously published multi-band phase shifters are high gain and digital control. Thus the measured input and output return losses are newly added for multi-bit applications.

In Chapter 2, the design, simulation, fabrication and performance of the dual-band band-pass/band-stop switching passive phase shifter are presented, showing that the dual-band 90-degree phase shifting becomes unavailable because of a low Q-factor of the circuit elements. In Chapter 3, the design, simulation, fabrication and performance of the dual-band band-pass/band-stop switching active phase shifter are presented, showing that the dual-band 90-degree phase shifting becomes available by utilizing 180-degree out of phase performance of the differential amplifier. Moreover, it is clearly shown as a comparative analysis that the dual-band active phase shifter based on band-pass and band-stop

*Corresponding Author: Yasushi Itoh, 1-1-25 Tsujido-Nishikaigan, Fujisawa, Kanagawa, 251-8511 Japan, +81-466-30-0185 | Email: itoh@elec.shonan-it.ac.jp

designs has an advanced feature in performance and controllability over the previously published multi-band phase shifters.

2. Dual-Band Band-Pass/Band-Stop Switching Passive Phase Shifter

2.1. Circuit Design

A schematic diagram of the dual-band band-pass/band-stop switching passive phase shifter is shown in Figure 1. It employs band-pass/band-stop filters comprised of dual-band resonators and SPDT switches. A 90-degree phase shifting can be accomplished by switching band-pass and band-stop filters.

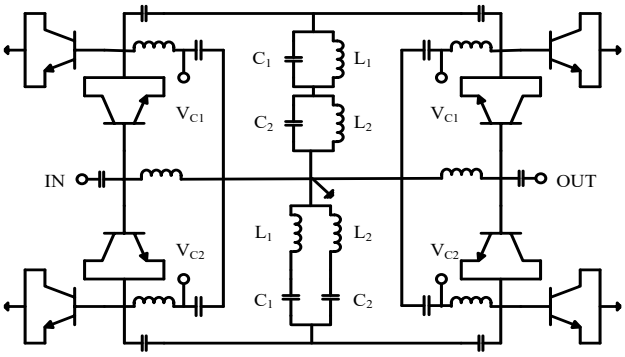


Figure 1: Schematic diagram of the dual-band band-pass/band-stop switching passive phase shifter

A dual-band phase shifting technique by switching band-pass and band-stop filters is illustrated in Figure 2. Figures 2(a) and 2(b) present band-stop and band-pass filters, respectively. The magnitude and phase shift of the band-stop and band-pass filters are illustrated in Figure 2(c) and Figure 2(d), respectively. The band-stop filter is comprised of two parallel LC circuits in a stacked form. It has a single band-stop frequency of f_{3a} and dual band-pass frequencies of f_1 and f_2, which are given as the following equations [11]-[12]:

$$f_1 = \frac{1}{2\pi\sqrt{L_1 C_1}} \tag{1}$$

$$f_2 = \frac{1}{2\pi\sqrt{L_2 C_2}} \tag{2}$$

$$f_{3a} = \frac{1}{2\pi}\sqrt{\frac{1/L_1 + 1/L_2}{C_1 + C_2}} \tag{3}$$

On the other hand, the band-pass filter is constructed from two series LC circuits. It has a single band-pass frequency of f_{3b} and dual band-stop frequencies of f_1 and f_2. f_{3b} is given as the following equation [11]-[12]:

$$f_{3b} = \frac{1}{2\pi}\sqrt{\frac{1/C_1 + 1/C_2}{L_1 + L_2}} \tag{4}$$

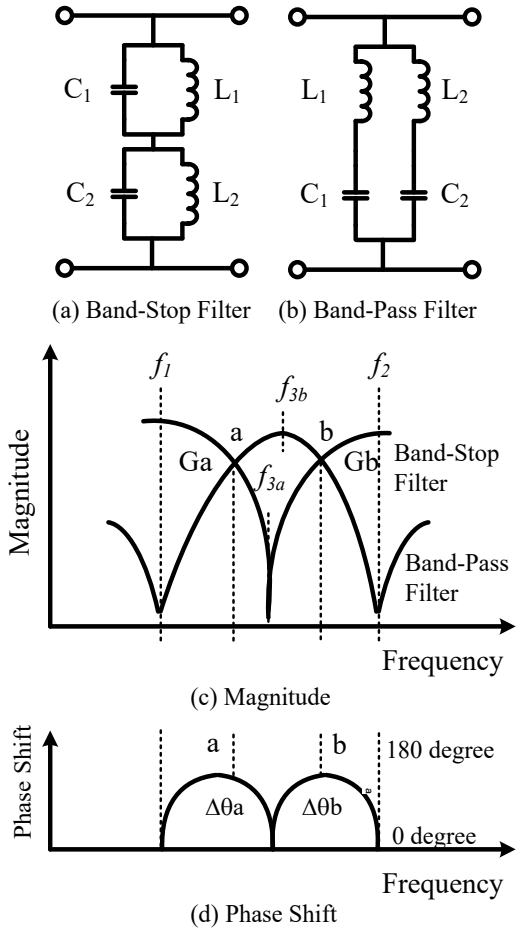


Figure 2: Dual-band phase shifting technique by switching band-pass and band-stop filters

At around transition frequencies in Figure 2(c), the impedance drastically changes from open to short, or short to open. Moreover, a low-pass filter shows a negative phase shift and a high-pass filter provides a positive phase shift. Due to the actual low Q factor, however, the phase shift obtained by switching band-pass/band-stop filters has a limit shown in Figure 2 (d). Now it is assumed that the gain is expressed as G_a, G_b and the phase shift is $\Delta\Phi_a$, $\Delta\Phi_b$ at the frequency points a or b, respectively. The circuit design has to focus mainly on how to determine L_1, C_1, L_2 and C_2 to meet with $G_a = G_b$ and/or $\Delta\Phi_a = \Delta\Phi_b$ at dual-band of the frequency points a and b.

2.2. Circuit Simulation

A circuit simulation is performed by using ADS2016 for the schematic diagram shown in Figure 1. 1005-type chip elements are used as R, L or C as well as 0.35 μm SiGe HBT with an f_t of around 25GHz is used as a switching element by taking into account the operating frequency around L-Band as well as better performance of SiGe HBTs over CMOS devices. The circuit element values are listed in Table 1. The simulated S_{21} and a phase shift ΔΦ is shown in Figure 3. At the frequency points a and b, ΔΦ becomes around 80 and 50 degrees, respectively

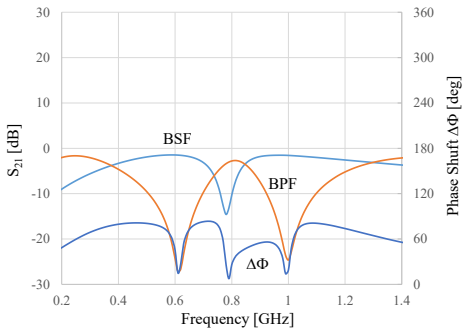


Figure 3: Simulated S_{21} and a phase shift ΔΦ

Table 1: Circuit element values

Element	Value
L_1 [nH]	8
L_2 [nH]	6.2
C_1 [pF]	8
C_2 [pF]	3

2.3. Circuit Fabrication and Performance

A photograph of the dual-band band-pass/band-stop switching passive phase shifter is shown in Figure 4. The phase shifter was fabricated on the FR-4 substrate with a dielectric constant of 4.4 @ 1 GHz and a tanδ of 0.016 @ 1 GHz. 1005-type chip resistors, capacitors, and inductors are mounted on the substrate by soldering. A surface mount type of the 0.35 μm SiGe HBT with an f_t of around 25 GHz (Toshiba MT4S102T) is used as a switching device. The circuit size is 16 x 16 x 1.2 mm^3.

The measured S_{21} and phase shift of the dual-band band-pass/band-stop passive phase shifter are demonstrated in Figure 5. The measured results are basically in good agreement with the simulated results of Figure 3. A slight discrepancy appears for the phase shifting value at G_a and G_b. This is most likely due to the error in the modeling of Q-factor. Bias conditions are V_{C1}=-1V and V_{C2}=+1V for BPF as well as V_{C1}=+1V and V_{C2}=-1V for BSF.

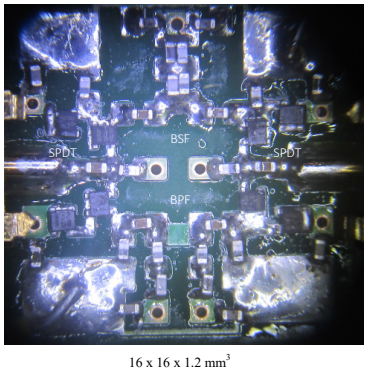


Figure 4: Photograph of the dual-band band-pass/band-stop switching passive phase shifter

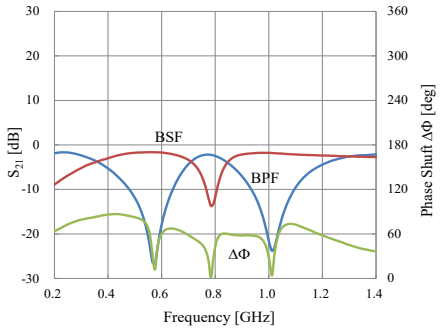


Figure 5: Measured S_{21} and phase shift of the dual-band band-pass/band-stop switching passive phase shifter

3. Dual-band Active Phase Shifter Based on Band-Pass and Band-Stop Designs Using Dual-Band Resonators

3.1. Circuit Design

A schematic diagram of the dual-band active phase shifter based on band-pass and band-stop designs is shown in Figure 6. It is comprised of a differential amplifier and a SPDT switch. The differential amplifier employs two different loads of Z_{L1} and Z_{L2}. Z_{L1} corresponds to the band-stop filter in Figure 2(a) and Z_{L2} in Figure 2(b). To supply a voltage to differential transistor pairs, a large inductance of L_3 (100nH) is connected in parallel with C_1 and C_2. Z_S is a series feedback circuit comprised of two resistors R_S. R_E is a current source. Differential outputs of OUT+ and OUT- are connected to a SPDT switch.

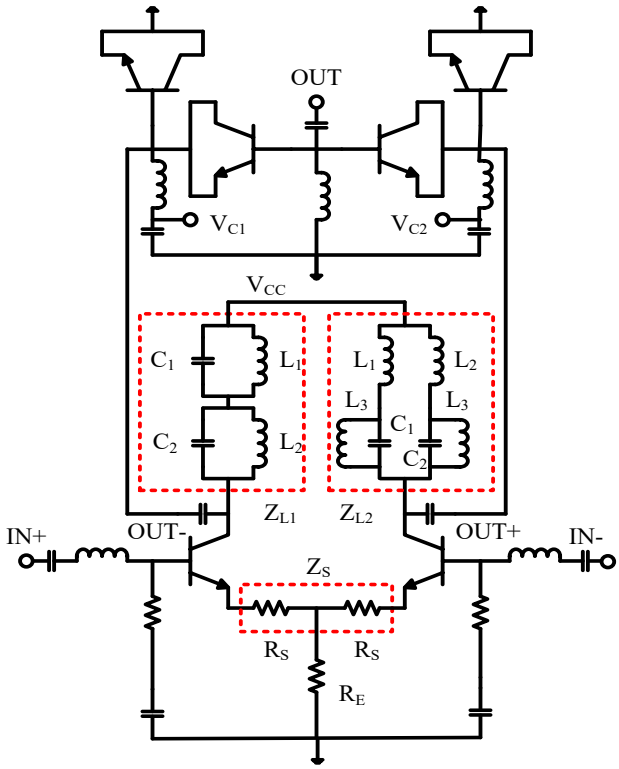


Figure 6: Schematic diagram of the differential amplifier and dual-resonant RLC circuit

The SPDT switch employs a series-shunt configuration for a high isolation. Since the collector and emitter are short-circuited, the transistor acts as a switching diode. V_{C1} and V_{C2} are a control voltage. Input matching circuits employ a lossy match configuration to improve an input matching as well as stabilize the circuit. The base bias resistors for a differential transistor pair are not shown in Figure 6 because of a circuit simplicity. V_{CC} is a supply voltage.

The phase shifting performances are compared for the passive phase shifter of Figure 1 and the active phase shifter of Figure 6, which is shown in Figure 7. Since the output signal from the differential amplifier is already 180 degrees out of phase, the phase shifting performance of the active phase shifter is a complementary value of the passive phase shifter.

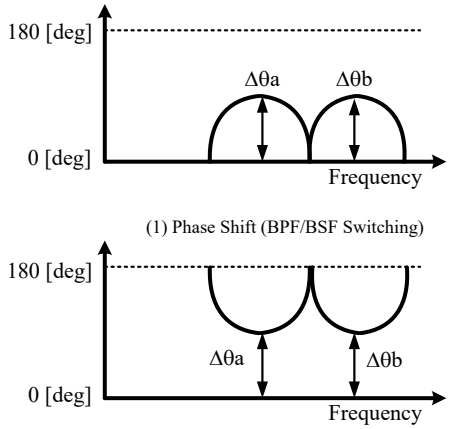


Figure 7 Phase shifting performances of the passive phase shifter (Figure 1 and the active phase shifter (Figure 6)

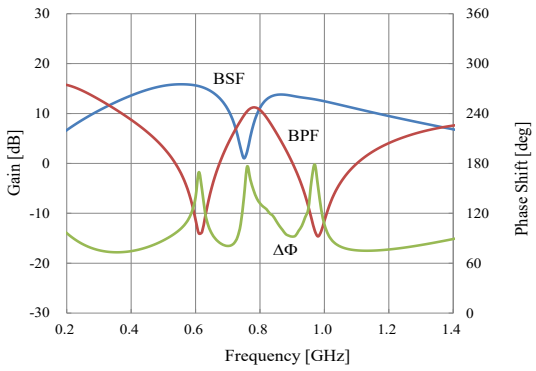


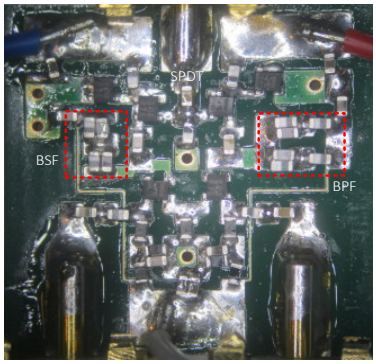
Figure 8: Simulated gain and phase shift of the dual-band active phase shifter using band-pass and band-stop designs

3.2. Circuit Simulation

The circuit simulation has been done by using ADS2016 for the schematic diagram in Figure 6 with the circuit element value of Table 1. V_{CC} is 6 V. V_{C1} and V_{C2} are +1 V or -1 V. Simulated gain and phase shift of the dual-band active phase shifter using band-pass and band-stop designs are shown in Figure 8. First, it is clearly shown that 180-degree phase shifting has been obtained at f_1, f_2, f_{3a} or f_{3a}. Moreover, a phase shift of 80 to 180 degrees can be realized between f_1 and f_{3a} or f_{3b} as well as between f_2 and f_{3a} or f_{3b}. It must be noted that the phase shifting value is greatly dependent on the frequency separation of f_1 and f_2 in addition to Q-factors of the band-pass and band-stop filters.

3.3. Circuit Fabrication and Performance

A photograph of the dual-band active phase shifter using band-pass and band-stop designs is shown in Figure 9. In a similar way as Figure 4, the active phase shifter was fabricated on the FR-4 substrate with a dielectric constant of 4.4 @ 1 GHz and a tanδ of 0.016 @ 1 GHz. 1005-type chip resistors, capacitors, and inductors are mounted on the substrate by soldering. A surface mount type of the 0.35 μm SiGe HBT with an f_t of around 25 GHz (Toshiba MT4S102T) is used as an amplifier device as well as a switching device. The circuit size is 16 x 16 x 1.2 mm^3. The circuit has two SMA connectors on the bottom side for differential inputs and a single SMA connector on the upper side for the output of the SPDT switch.



16 x 16 x 1.2 mm^3

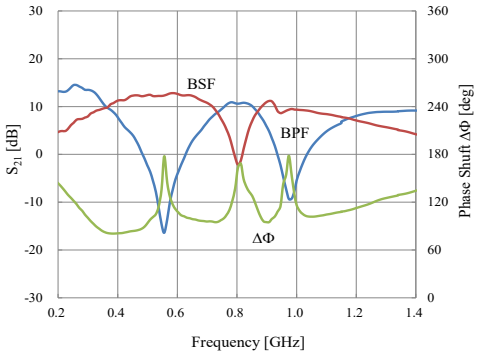
Figure 9: Photograph of the dual-band active phase shifter using band-pass and band-stop designs

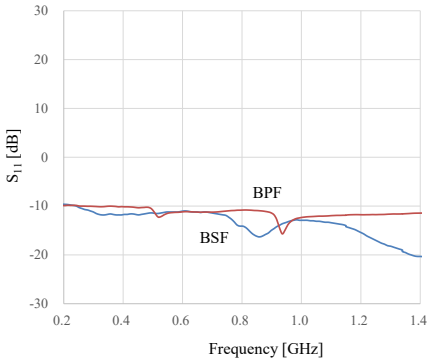
The measured gain and phase shift of the dual-band active phase shifter using band-pass and band-stop designs using dual-band resonators are demonstrated in Figure 10. The measured results are basically in good agreement with the simulated results of Figure 8. The active phase shifter has achieved a gain of 9.3 dB and a phase shift of 95 degrees at 0.74 GHz as well as a gain of 8.9 dB and a phase shift of 98 degrees at 0.88 GHz. Small gain ripples appear in Figure 10. This is most likely due to the stability problem. Although the stability problem can be improved with the use of the load resisters, the phase shifting amount becomes smaller due to a low Q-factor, which provides a design tradeoff. Bias conditions are V_{CC}=6 V and V_{C1}, V_{C2}=+1V or -1V, respectively.

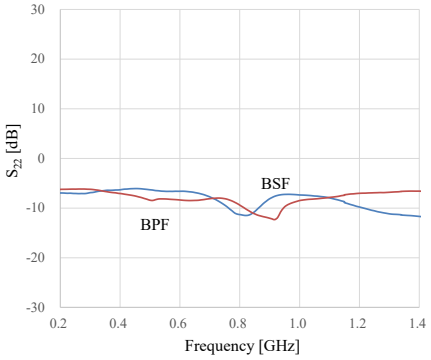
The measured input and output return losses of the dual-band active phase shifter using band-pass and band-stop designs are demonstrated in Figures 11 and 12, respectively. The input return loss was 11.8 dB at 0.74 GHz and 16.3 dB at 0.88 GHz. The output return loss was better than 8 dB at both 0.74 GHz at 0.88 GHz. As mentioned above, the load resistor also improves the output return loss but the Q-factor becomes worse.

Table 2: Comparative analysis

References	Year	Application	Phase Shifter Type	Dual-Band Method	Active/Passive	Controllability	Frequency	Phase Shifting	Insertion Gain	Return Loss
[4]	2006	WLAN	Reflection	Switching Varactor Diodes	Passive	Analog	2.4/5.2 GHz	120°/180°	-3.0/-5.0 dB	-15/-10 dB
[6]	2011	Doherty PA	Loaded-Line	Switching Capacitive Loads	Passive	Analog	2.1/2.2 GHz	90°	not shown	not shown
[8]	2004	WLAN	Vector-Sum	Switching Phase Shifters	Active	Analog	2.4/5.2 GHz	360°	+4.0/+3.5 dB	not shown
This Work	2018	under R&D	BPF/BSF	Switching BPF/BSF	Active	Digital	0.74/0.88 GHz	90°	+9.3/+8.8 dB	-8 dB







3.4. Comparative Analysis

A comparative analysis is summarized in Table 2. Since this paper is the first report on the dual-band phase shifter using BPF/BSF or LPF/HPF switching, the results obtained from this work are compared with the other multi-band phase shifting methods. All of the multi-band phase shifters in Table 2 have achieved a multi-band phase shifting by switching devices or circuits. The reflection type [4] and the loaded-line type [6] are based on passive phase shifter designs and thus the insertion loss is large. On the other hand, the vector-sum type [8] and our work [1] are an active phase shifter and thus have an insertion gain. Moreover, the phase shifters in [4], [6] and [8] are analogously controlled and thus the complex control is needed. It can be concluded from these viewpoints that the dual-band phase shifter using band-pass and band-stop designs of our work can provide several advantages especially in performances and controllability over the previous ones.

4. Conclusions

Design, fabrication and performance of the dual-band active phase shifter based on band-pass and band-stop designs using dual-band resonators have been presented. It can be clearly shown that the dual-band 90-degree phase shifting is unavailable for the passive phase shifter merely using band-pass and bandstop switching but becomes available for the active phase shifter using a differential amplification. The implemented dual-band active phase shifter using 0.35 μm SiGe HBT has achieved a gain of 9.3 dB and a phase shift of 95 degrees at 0.74 GHz as well as a gain of 8.9 dB and a phase shift of 98 degrees at 0.88 GHz. 180-degree phase shift can be easily obtained with the use of the differential amplifier, In addition, a lower bit of 45-, 22.5- and 11.25-degrees can be achieved by using the same band-pass and band-stop designs. With the use of these techniques, digitally-controlled 5-bit multi-band active phase shifters having a high gain can be realized for use in the next generation, multi-band and multi-mode wireless communication systems.

References

[1] Y. Itoh, H. Takagi, “A Dual-Band 90-Degree SiGe HBT Active Phase Shifter Using Band-Pass and Band-Stop Designs”, in 47th European Microwave Conference, pp. 232-235, October 2017.

[2] T. Nakagawa, M. Kawashima, H. Hayashi, K. Araki, “A 0.9-2.5 GHz Wideband Direct Conversion Receiver", in IEEE GaAs IC Digest, pp. 37-40, 2001.

[3] H. Kamizuma, T. Yamawaki, Y. Akamine, K. Maeda, S. Tanaka, “A Quad Band WCDMA Transceiver with Fractional local divider", in Symposium on VLSI Circuits Digest, pp. 96-99, 2008.

[4] Ocera, E. Sbarra, R. Vincenti Gatti, R. Sorrentino, “An Innovative Reconfigurable Reflection-Type Phase Shifter for Dual Band WLAN Applications", in European Microwave Association, pp. 64-67, 2006.

[5] Y. Shoji, K. Sakamoto, Y. Itoh, “A Dual-Band Reflection Type Phase Shifter Using Active Loads”, in Contemporary Engineering Sciences, Vol. 7, No. 20, pp. 957-964, 2014.

[6] N. Park, J. Gwan, Y. Kim, C. Lee, "Dual-band Switching Doherty Power Amplifier using Phase Shifter Composed of PIN Diode", in 6th European Microwave Integrated Conference, pp. 300-3003, 2011.
[7] M. Maassel, B. D. Braaten, D. A. Rogers, "A Metamaterial-Based Multiband Phase Shifter", in IEEE International Conference on Electro/Information Technology, pp. 533-535, 2014.
[8] D. R. Banbury1, N. Fayyaz2, S. Safavi-Naeini1, S. Nikneshan, "A CMOS 5.5/2.4 GHz Dual-band Smart-antenna Transceiver with a Novel RF Dual-band Phase Shifter for WLAN 802.11a/b/g", in IEEE RFIC Symposium Digest, pp. 157-160, 2004.
[9] Lu, A. Pham, D. Livezey, "On the Linearity of CMOS Multi-band Phase Shifters", in Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems, Page 4, pp., DOI: 10.1109 /SMIC.2005.1587973, 2006.
[10] Z. El-Khatib, L. MacEachern, S. A. Mahmoud, "Fully-Integrated Multi-band Tunable Linearized CMOS Active Analog Phase Shifter with Active Loss Compensation for Multiple Antenna Wireless Transceiver Applications", in International Symposium on Circuits and Systems, pp. 2966-2969, 2009.
[11] G. L. Matthaei, L. Young, E.M.T. Jones, Microwave Filters, Impedance-Matching Networks, and Coupling Structures, Artech House, ISBN: 978-0-89006-099-5, 1965.
[12] M. Poger, "Microwave engineering education: From field theory to circuit theory", in IEEE/MTT-S International Microwave Symposium Digest, 2012.
[13] Y. Ayasli, S. W. Miller, R. Mozzi, L. K. Hanes, "Wide-Band Monolithic Phase Shifter", in IEEE Trans. ED-31, No. 12, pp. 1943-1947, 1984.
[14] J. P. Comeau, M. A. Morton, J. D. Cressler, J. Papapolymerou, M. Mitchell, "A High-Linearity 5-bit, X-band SiGe HBT Phase Shifter", in IEEE MTT-S Digest, pp. 1668-1671, 2006.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 226-234 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Contactless Power and Bidirectional Data Transmission via Magnetic Field

Jia-Jing Kao, Chun-Liang Lin*, Chih-Cheng Huang, Hau-Shian Jian

Department of Electrical Engineering, National Chung Hsing University, Taichung, Taiwan

A R T I C L E I N F O

Article history:
Received: 24 June, 2018
Accepted: 05 September, 2018
Online: 05 October, 2018

Keywords:
Wireless power transfer method
Electric vehicle charging
Magnetic field transmission
ZVS

A B S T R A C T

At present, the wireless charging and data transmission systems developed by major car manufacturers rely heavily on additional transmission equipment. This paper presents a newly developed wireless power and bidirectional data transmission scheme without adding any radio frequency (RF) devices. The battery charge status and vehicle related information can be transmitted bidirectionally in the wireless manner by the compensation capacitor between two isolated units. Based on the function, this system can quickly inform the user when an emergency event occurs during power transmission. An inverter is used on the primary side of the system, and the generated AC power is transmitted to the load at the secondary side through mutual inductance; the secondary side of the system adjusts the load current to transmit data back. The primary side uses the zero voltage switching (ZVS) method to receive data, and the trimming of the current curve to transmit commands; the secondary side of the system receives commands and uses carrier cycles for decoding. Experimental verification shows applicability of the proposed system.

1. Introduction

The principle of inductive power transfer (IPT) is the mutual magnetic coupling between the transmission windings. It transfers the main power source on the fixed side to a plurality of movable devices by means of the air gap [1]. The issue is becoming popular due to the increasingly global concern of energy conservation. However, most of the previous research tasks only focus on the issue of IPT without considering simultaneous power and data transmission. This research intends to deal with the issue with a novel idea. The paper presented here is an extension of our work originally presented in the *2017 12th IEEE International Conference on Industrial Electronics and Applications* [2] by including extended results.

Traditional transferring power is accomplished through the contact of the metal contacts (Plug-In). Although this kind of power transmission can meet most of the requirements, it is inconvenient to be used in many occasions, such as moisture, snow, oil exploitation, and underwater power supply. In order to avoid the shortcomings of contact power transmission, contactless power transfer (CPT) has been developed, which is one of the popular wireless power transfer (WPT) technologies. It has no metal contact and hence avoiding the weakness.

The wireless charging method has been receiving attention over a wide range of industrial and civilian applications in recent years which leads to a variety of utilizations such as contactless electric vehicle charging [3-5], bidirectional IPT system [6-8], electric vehicle and power transmission between grid [9], wireless power and data transmission with fully implantable stimulator [10], portable-telephone charging [11], dynamic charging of electric vehicle [12-13], and wireless power and data transmission system [14].

In literature [7], the proposed IPT system has been applied to vehicle battery charging with high efficiency. Its control loop is composed of external and internal paths, used to handle power control between two sides with an additional wireless transmission equipment to transfer data from the secondary side to the primary side. This design requires better transmission security between two isolation units. The IPT system of [15] uses non-contact technology in the drilling machine, and two sets of windings are used for power supply and data transmission respectively. The design would lead to extra production cost.

The architecture of this research task is the use of a flyback converter, which is used for wireless power and bidirectional data transmission, and adopts ZVS technology to reduce switching loss and EMI noise. The secondary side of the system performs data modulation by instantaneously adjusting load current, and the primary side of the system receives data under ZVS. In addition, the system trims the current transmitting command on the primary side, and the secondary side of the system receives command and decodes data. The authors of [15] proposed a transient load detection method, which mainly uses energy injection and free resonance mode to detect the load. Although a similar approach was used here, wireless power and bidirectional data transmission

* Chun-Liang Lin, Email: chunlin@dragon.nchu.edu.tw

are the focus of this research. Our design proposes an effective power and bidirectional data transmission solution under the architecture of the IPT system. The technology can be combined with a vehicle positioning system to realize a wireless vehicle charging station as the final goal. As for a practical application in the micro grid, when a driver is charging his/her car, if the charge station can immediately know the voltage, current, power and related information of the charging car, the accident can be effectively managed. Beside of this, if the car is used as a mini power generator, the grid side may identify car ID information and count how much power it contributing to the grid.

2. Description of System

2.1. Circuit Design

This research task considers a wireless vehicle charging system with the function of synchronous power and bidirectional data transmission. The primary side (Grid side) uses a utility power supply (AC 110V / 60Hz). The secondary side (Vehicle side) stores the energy transmitted from the grid side in the vehicle battery. In addition to power transmission, both sides can send and receive messages simultaneously through the same conduction windings. At startup, the primary side of the system performs two steps in sequence, namely payload detection and soft start. First, it detects whether there is a payload on the secondary side of the system. If no payload is detected on the pad's inductance, the primary side stops power output. If a load is detected, the soft start current will give rise to a preset level, and this mechanism minimizes the inrush current flowing through the switch. The schematic diagram of the proposed system is illustrated in Fig. 1.

The flyback converter adopted here is widely used in AC-DC conversion or DC-DC conversion. Its operational principle is similar to the buck-boost converter. In addition to voltage conversion, there is electrical isolation between the input and output circuits.

The system consists of two circuits that are magnetically coupled to each other. The grid-powered battery charging station (GPBCS) converts the main power supply to direct current through a rectifier and a filter. The transmitter generates power at the primary windings and transmits it to the secondary windings, which is located inside of the vehicle, by mutual inductance to complete the wireless power transfer.

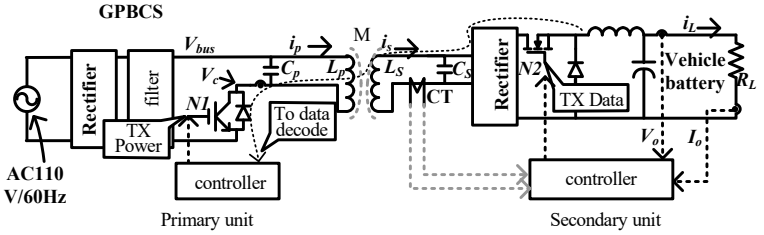
The power switch (MOSFET) N_2 at the secondary side has two functions. First, it bypasses the output current and then sends "T_X data" such as output voltage and current back to the primary side. The way in which the primary side receives data is to detect the period of the feedback voltage V_C. For data communication, when the T_X data is received and confirmed by the primary side, the switch (N_1) at the primary side sends a command to the secondary side. The current transducer (CT) at the secondary side is connected to the circuit to receive the command.

As can be seen from Fig. 1, the magnetic coupling coefficient and the load quality factor Q of the secondary side are closely related to the mutual inductance M. Its equation is as follows:

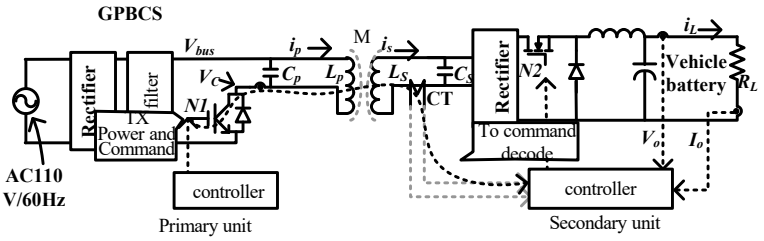
$$k = \frac{M}{\sqrt{L_p L_s}}$$

$$Q = \omega C_s R_L$$

where L_P and L_S are primary side inductance and secondary side inductance, respectively and C_S is secondary side capacitance, R_L is load on the secondary side. The operating frequency and rated power of the power supply are mainly limited to the switching components used, both must be balanced according to the copper losses. In this research, the appropriate Q and k were chosen to maintain efficiency of power transmission [5], and S/N ratio, thereby controlling quality of the transmitted data.



(a) Transmission path of data feedback to the primary (Grid) unit



(b) Transmission path of command sending to the secondary (Vehicle) unit

Figure 1. Simultaneous wireless power and bidirectional data transmission scheme

The simultaneous wireless power and bidirectional data transfer system adopts parallel–parallel (PP) reactive power compensation topology. The primary side for parallel compensation is used to generate a large primary current; the secondary side suitable for battery charging current sources [16]. The PP type compensation topology is illustrated in Fig. 2 (a). The topology under mutual inductance coupling can be represented by the simplified circuit depicted in Fig. 2 (b).

As shown in Fig. 2, the impedance reflected from the secondary side back to the primary side is given by

$$Z_r = \frac{\omega^2 M^2}{Z_s} \tag{3}$$

The voltages of the primary windings and secondary windings are expressed as follows:

$$V_p = -j\,\omega M I_s + j\,\omega L_p I_p$$

and

$$V_s = -j\,\omega L_s I_s + j\,\omega M I_p$$

The current of the secondary winding can be expressed by the following equation:

$$I_s = \frac{j\,\omega M I_p}{Z_s}$$

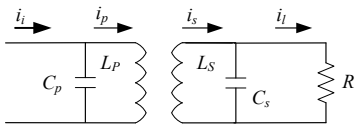
with the resonant frequency of both sides being

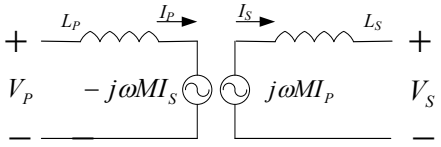
$$\omega = \frac{1}{\sqrt{L_p C_p}} = \frac{1}{\sqrt{L_s C_s}}$$

The secondary side also transmits data to the primary side with the same mutual inductance. The load impedance of the primary side can be known for the network combination of the primary side and the secondary side, namely,

$$Z_t = \frac{1}{j\omega C_p + \frac{1}{j\omega L_p + Z_r}}$$

$$C_p = \frac{\left(L_p L_s - M^2\right) C_s L_s^3}{M^4 C_s R + L_s \left(L_p L_s - M^2\right)^2} \tag{9}$$





(b) Mutual inductance coupling circuit

Figure 2. Compensation topology and the simplified circuit

To minimize the VA rating, the imaginary part of the load impedance is adjusted to zero when the secondary side is working at the resonant frequency. The selected primary side capacitance C_P is as follows

3. Wireless Data Transmission

3.1. Communictaion Protocol

The data packet can be divided into four parts, namely start segment, data segment, checksum segment, and stop segment. The entire packet architecture is referenced to the 1-Wire communication protocol and fine-tuned to construct a packet suitable for use with the system. In addition, the universal asynchronous receiver transmitter (UART) specific start segment and stop segment are also retained in the system to separate the packets. The major data types in the data segment are command, voltage value, current value, power, and stop power transmission. The checksum segment is used to detect the entire packet error.

The messages we are about to transmit are divided into two categories based on their attributes. The first category is normal messages including output voltage and current, which is handled with a slower rate. The second category is emergency messages such as overvoltage and overcurrent, which is processed with a faster rate for instantaneous reaction by the system. In addition, when T_X data of the secondary side is sent back to the primary side, the system will calculate the charging efficiency. The primary side is the command issuer, which will send the T_X cmd to the secondary side. Table I summarizes the specifications for data transmission.

Table 1: Protocol for data transmission

Category	Baud rate	Start segment	Data segment	Checksum segment	Stop segment
Normal messages	120 Bd.	1 bit	8 bits	8 bits	1 bit
Emergency messages	Operating freq. Bd.	1 bit	4 bits	4 bits	1bit

3.2. Data Transmission

When the primary side transmits power to the secondary side, description of data transmission between two sides is explained as in Fig. 3. In this figure, the period between zero-crossing will coincide with the AC line. The time base clock on the primary side is the “Digital AC line” signal generated by the zero-crossing detection circuit. In order to be synchronized with the primary side of the Digital AC line, the secondary side uses a current sensor to measure I_S, which is then compared to produce a “Digital I_S” signal.

Signal modulation technique is commonly adopted in communication and electronic systems to mix and transmit signals for data transmission in one or more carrier waves. General modulation modes include amplitude-shift keying (ASK), frequency-shift keying (FSK), and phase-shift keying (PSK). In which, FSK has been widely used in digital signal modulation for communication systems such as emergency broadcasts. It uses discrete frequency changes of the carrier to modulate data for transmission. In which, the carry frequency f_C is used to

discriminate digital "0" or "1". Referring to Fig. 3, in the preamble section, T_{OFF} of the duty cycle of N_2 is fixed. When the secondary side transmits data in the data section, the primary side receives the data in the shadow section; when the primary side transmits the command in the cmd section, the secondary side receives the command in the shadow section.

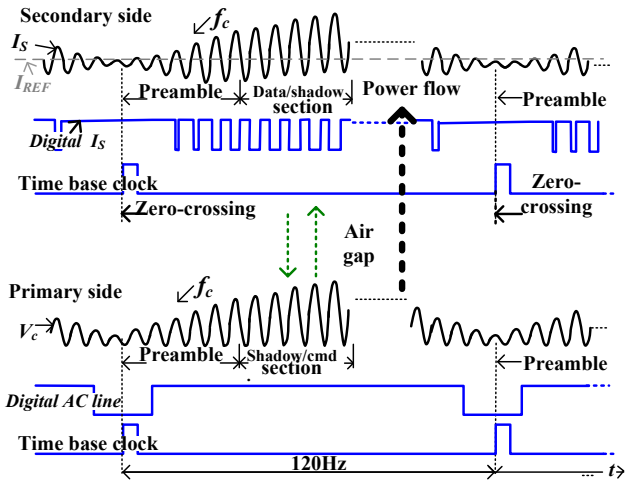


Figure 3. Description of bidirectional data transmission via the mutual magnetic coupling circuits

3.3. Data Decoding

When data is allocated in the data section on the secondary side and modulated to $T_Z(0)$, the controller reduces the duty cycle of N_2 and the output current (See Fig. 1). The reduced current causes a time delay on the secondary side, which we denote it as $\Delta\tau$. Further, as can be seen from Fig. 4, N_2 is turned off in the c-d section so that the current I_S flows through the load R_L is stored in C_S; when N_2 works in the d-e section, it is turned on, and C_S will discharge to R_L. The delayed discharge of C_S directly affects V_C on the primary side, causing V_C to generate a delay time at 0V. At this time, the d-e section corresponding to V_C will have a bump. Since N_1 operates based on ZVS, the turn on time of N_1 is also delayed. The additional delay time produced by the a-e section is also defined as $\Delta\tau$ after comparison with the e-g section. The time delay $\Delta\tau$ resulting from modulation of the digit "0" is as follows

$$T_Z(0) = T_C(1) + \Delta\tau \tag{10}$$

where $T_Z(0)$ under modulation denotes the period of the carrier wave for digit "0" and $T_C(1)$ for the digit "1". The term $\Delta\tau$ can be referred to the S/N ratio. An increase in $\Delta\tau$ will result in better data transmission quality, making it more robust against noise, and it is more advantageous for the controller on the primary side to decode the signal. However, it might deteriorate the power transmission quality.

It can be known from the above that the secondary side can send the numbers "0" and "1" by $T_Z(0)$ and $T_C(1)$ respectively in the data section; the primary side receives the modulated data, which being decoded as "0" or "1" is based on $T_Z'(0)$ or $T_C(1)$. The rule for the primary side distinguishes the digital "0" is determined by

$$T_Z'(0) > T_C(1) + \Delta ref \tag{11}$$

where Δref is a threshold set to identify the digit "0" at the primary side. Illustration of the idea is depicted in Fig. 5 which explains the function for modulating a digit "0" at the center of the data string where curve 1 is I_S, curve 2 is the switching state of N_2, curves 3 and 4 are the gate voltages of V_C and N_1, respectively. After changing the duty cycle of N_2 (100% adjusted to 94%), it can be seen that the waveform of I_S becomes the defined $T_Z(0)$; the waveform of the synchronous V_C also becomes the defined $T_Z'(0)$. As can be seen from the results in the figure, when the periods of I_S and I_P are the same, the operating frequency will be affected when I_S changes.

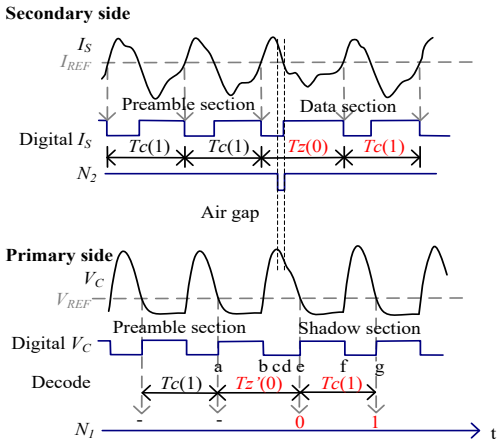


Figure 4. Data decoding method

Fig. 6 shows the varying duty cycle of N_2 vs. $\Delta\tau$. It can be seen from the figure that after adjusting the duty cycle, a near linear relationship can be obtained between 85% and 94%. The relationship is as follows:

$$\text{Modulated } T_{off} \approx \Delta\tau + \varepsilon \tag{12}$$

where the modulated T_{off} is duty off of N_2 for the secondary side to send the digit "0", the uncertainty $|\varepsilon| \le 0.5\mu s$ is set in our experiments. If the modulated $T_{off} + \varepsilon > \Delta ref$, the primary side decodes data as the digit "0".

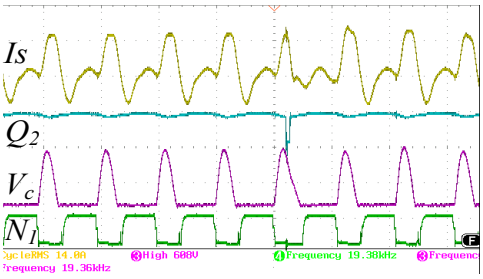


Figure 5. The digit "0" is modulated by reducing the duty ratio of N_2 from 100% to 94% when transmission power is at 700 W

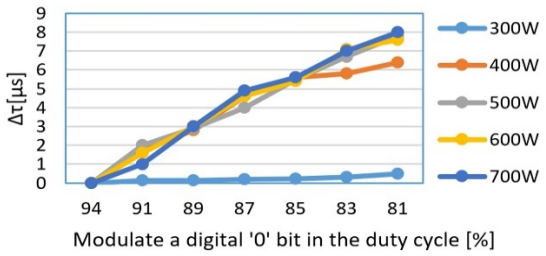


Figure 6. Relationship between duty cycle of N_2 and $\Delta\tau$

3.4. Command Transmission

In order to prevent conflicts during data transmission, the system-defined data arrangement string is shown in Fig. 7 where T_X cmd is the corresponding command message generated on the secondary side after sending the message on the primary side. The reset time is set to wait for the next string. As for T_X data, it is the feedback from the normal message of the secondary side, such as output voltage and current. Fig. 8 depicts how the system processes T_X cmd synchronously during the wireless power transfer.

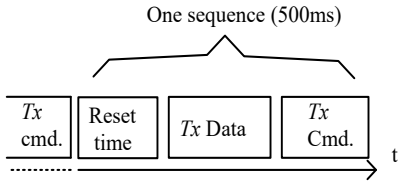


Figure 7. Data transmission scheme

When the system is operating normally and there is no emergency on the secondary side, the primary side places command in the Cmd. section with baud rate 120 Bd. To react, the secondary side detects command accordingly in the shadow section from "Digital I_S". When modulation is performed to the digital "0", the primary side changes the duty cycle of the switch N_1. As seen from Fig. 9, the time required for T_{ON} is shorter when modulating "0" and the time required for T_{ON} when modulating "1" is longer. The reduced current causes a loss of time on the primary side, which is what we call $\Delta\tau$. Referring to Fig. 8, the primary side will arrange data to be sent in the Cmd. section. The turn on time N_1 in the b-c section will be shorter than other parts. The current change of I_P will affect the secondary side, which is to be explained later. It causes a reduced time period $\Delta\tau$ at the shadow part in the a-d section than that in the d-e section. It can be found from the upper sub-figure of Fig. 8 that the term $\Delta\tau$ generating the digital messages "0" is defined as

$$T_Z^{P2S}(0) = T_C(1) - \Delta\tau \tag{13}$$

where $T_Z^{P2S}(0)$ denotes the primary side modulating a digit "0" to the secondary side. The secondary side receives command and decodes it to be "0" or "1" depending on T_C (1) and $\Delta\tau$. The secondary side distinguishes the digital "0" based on the following criterion:

$$T'^{P2S}_Z(0) < T_C(1) - \Delta ref \tag{14}$$

where $T'^{P2S}_Z(0)$ denotes the secondary side receiving a digit "0" and Δref is a threshold for judging the digit "0" at the secondary side. If $\Delta\tau > \Delta ref$, the primary side trims the duty cycle of N_1 as

$$\text{Trimmed } T_{on} \square\ \Delta\tau + \varepsilon \tag{15}$$

where $\Delta\tau$ and ε are defined as above. If Trimmed $T_{on} > \Delta ref$, the secondary side decodes data as the digit "0". Figs. 10 and 11 display the function for modulating a digit "0" ($\Delta\tau = 4\mu s$) at the center of the data string. It is seen that $T_Z^{P2S}(0)$, by modulating the duty cycle of N_1, is relatively shorter than that of modulating a "1" by $4\mu s$. Fig. 9 shows simultaneous wireless 700 W power transfer while simultaneously transferring *Tx* cmd the under the air gap of 25 mm and 0 mm lateral offset.

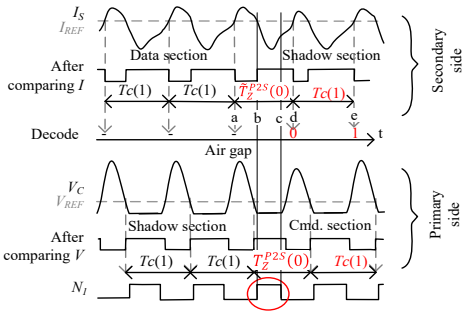


Figure 8. Working mechanism of command transmission

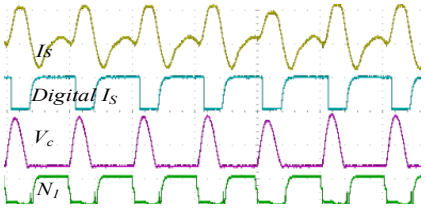


Figure 9. Simultaneous transmission of T_X cmd at 700 W power transmission (Under 25 mm air gap and 0 mm offset)

3.5. Emergency Message

If there is abnormal voltage or current detected during power transmission, the system will stop charging with the highest priority.

We reduce the duty cycle of N_2 from 100% to 94% to generate an additional delay time $\Delta\tau = 4\mu s$ and to confirm the result. The results are shown in Figs. 10 and 11 for the cases of 400 W or 700 W power transmission. It is seen that the primary side can still correctly decode data whether the IPT system works at low or high power.

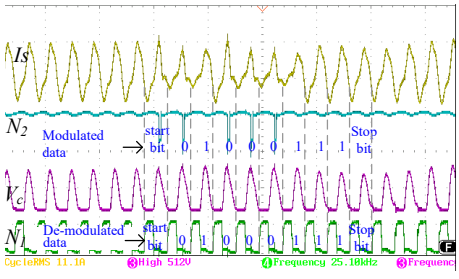


Figure 10. Synchronous transmission of emergency message when transmitting power at 400 W

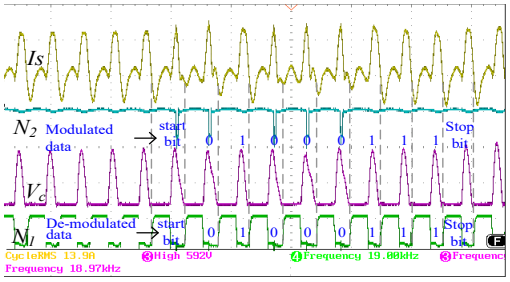


Figure 11. Synchronous transmission of emergency message when transmitting power at 700 W

3.6. Transmission Windings

IPT is a kind of electricity conversion technology. For operation, the primary side windings pass the alternating current I_P and the alternating magnetic field intensity H is generated which generates the magnetic flux density ψ via the air medium coupling. This affects the secondary side windings and induces the electromotive force ε (see Fig. 12).

The windings used in this research are the spiral type and the winding ends are not connected. In the real-world experiment, in order to reduce the internal impedance of the windings, the coil has a wire diameter of 2.0 mm to increase the efficiency of transmission. The diameter of the windings is 165 mm, the resonant frequency of the windings is 22 kHz when loaded, and the resonant frequency of the windings is 32 kHz without load.

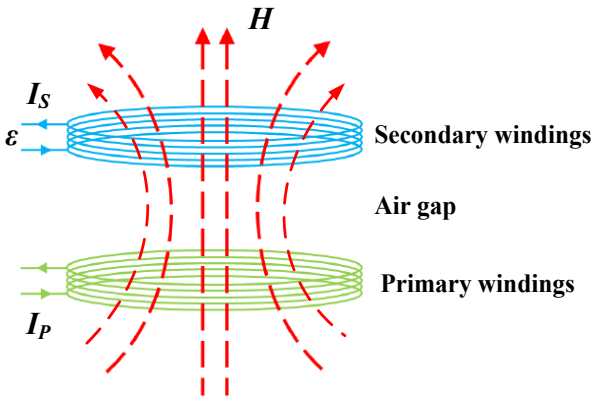


Figure 12. Electromagnetic induction diagram of two transmission windings

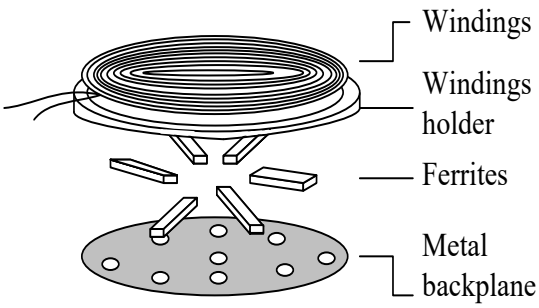


Figure 13. Transmission windings construction used in this research

3.7. Interface

The human machine interface (HMI) control circuit is shown in Fig. 14. It uses the Arduino MEGA 560 as the coordination hub of the entire HMI control circuit, and equipped with an HD44780 universal 16-inch liquid crystal display (LCD) to display the transmission information. The user control interface is implemented by a simple button circuit to facilitate the status display of bidirectional power and data transmission. It communicates with the MCU of the IPT system through a UART communication interface.

The Arduino MEGA 2560 receives the decoded data packet in real time. It also detects the loop of the button instantly. When the user presses the button, it sends a data packet to the IPT system for use by the encoding circuit. Fig. 15 shows the entity of the HMI.

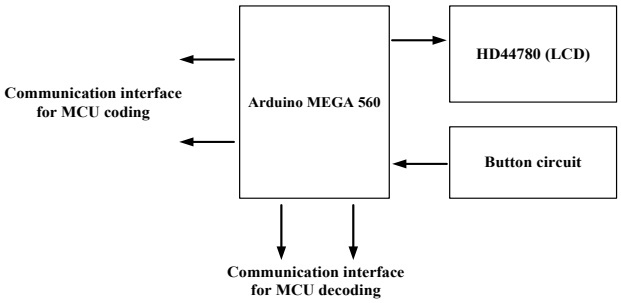


Figure 14. HMI diagram

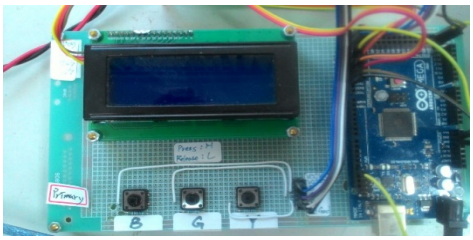


Figure 15. HMI circuit

3.8. Measurement

The primary side controls output power according to the demand of the secondary side. The later simultaneously receives power and measured voltage, current, and power via the way of magnetic field coupling.

The secondary side employs a voltage attenuator and a voltage amplifier to detect charging voltage and current. The crest factor of both signals was designed to equal to 3. These normalized analog signals were connected to a 12-bit ADC device shown as in Fig. 16. In the figure, in addition to measurement, the tasks of the CPU include data encoding, data decoding, and HMI process. To reduce power consumption, the ADC device utilizes the under-sampling method [17-18] to sample signals. It adopts low sampling rate and long measurement period to improve measurement accuracy. The measured raw voltage and current at the sampling time are given by

$$v_raw_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N} v_Gain\left(v_j - v_dc_offset\right)^2} \quad (16)$$

$$i_raw_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N} i_Gain\left(i_j - i_dc_offset\right)^2} \quad (17)$$

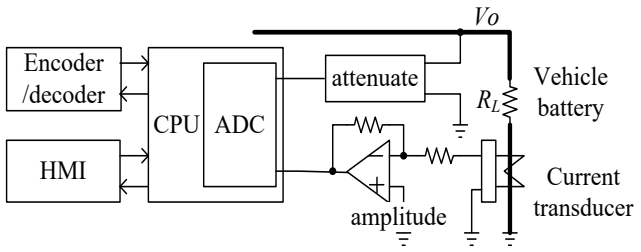


Fig. 16. The secondary control unit with measurement

where N is data update rate, v_Gain and i_Gain are voltage and current gain, respectively. These gains were generated by full-scale calibration. The terms v_j and i_j are the *jth* voltage and current sampling values, respectively, v_dc_offset and i_dc_offset are inherent dc-offsets.

Voltage and current are obtained by calculating

$$AVG(v_{raw_{i-K-1}} + v_{raw_{i-K-2}} + .. + v_{raw_{i-1}} + v_{raw_i}) \quad (18)$$

$$I_i = AVG(i_{i-K-1} + i_{raw_{i-K-2}} + .. + i_{raw_{i-1}} + i_{raw_i}) \quad (19)$$

where AVG denotes "average" and K is index of measurement period.

We use the under-sampling approach to measure charging voltage, current, and power. Comparison of the parameters of under-sampling and over-sampling methods is given in Table II. The experimental results are shown in Fig. 17. The measurement error is larger at low output power because of the nonlinear distortion of voltage attenuator and amplifier.

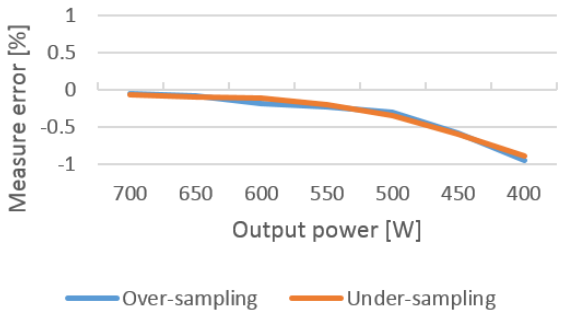


Figure 17. Comparison of the under-sampling and over-sampling methods

Table 2: Measured parameters of under sampling method vs. over-sampling

Parameter	Sampling type	
	Under sampling	Over sampling
ADC resolution [bits]	12	
Crest factor	3	
Data update rate [ms]	250	250
Sampling rate [kHz]	≒5	200
Measurement period [ms]	2000	250

3.9. Efficiency

The primary side controls output power based on the ZVS method. The operational frequency was varied by various output power or demand between 19~30 kHz. The sampling rate was 200 kHz. The channels of sampling voltage and current are V_{bus} and i_P, respectively. Their crest factor is 3. Both signals are connected to the input of ADC and controller for measurement.

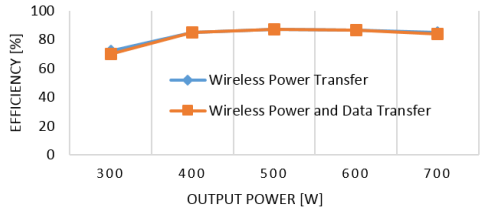


Figure 18. Comparison of efficiency change with output power

Drop of the power transfer efficiency is shown in Fig. 18 when power and data are transmitted simultaneously. As seen from the figure, this will only result in a slight decrease in the power transfer efficiency.

4. Experimental Results

Since the primary side operates at ZVS, the operating frequency of the system can be changed by the load R_L. In addition, Q_1 on the primary side can have less switching loss and higher efficiency under ZVS operation. If the air gap of two transmission windings changes, the operating frequency of the system will vary. To resolve solve the problem, both units measure the period of T_C (1) in the preamble section as a reference.

We conduct the following experiments to verify robustness of data transmission quality (TQ) while there is misalignment between two induction pads or there is load change. We consider Δref to be $\Delta\tau / 2$ and parameters listed in Table III. Fig. 19 shows the system hardware and test environment. Power transmission in both sides are controlled by ARM Cortex-M0. The clock speed of the microcontroller is 24 MHz. The data transmission quality (TQ) is defined by

$$CQ = \frac{pass\ count}{total\ count}\% \quad (20)$$

When transmitting a normal message, the total variable count is 1,000. The pass count for the number of correct packets is defined by the following equation:

$$CS = no.\ of\ Data\ 2's \quad (21)$$

Data received is treated as a correct message if the checksum (CS) is equal to *Data2's.*

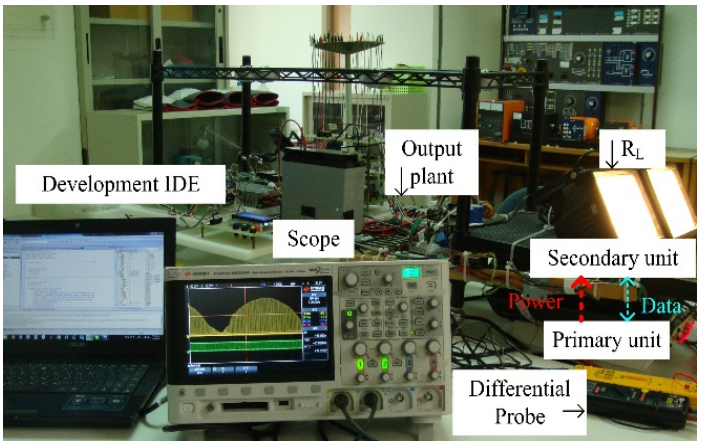


Figure 19. Hardware relaization of the IPT system

Fig. 20 presents the result of data TQ. It was recorded by adjusting the lateral shift of two pad's center from -20 mm to +20 mm. It is found from the experimental results that regardless of the lateral shift between the two pads, an increase in $\Delta\tau$ can significantly improve the data TQ. In addition, when the lateral shift of the two pads is maintained within -5 mm to +5 mm, and $\Delta\tau$ is set to between $3\mu s$ to $5\mu s$, the data TQ can be as high as 90% or more.

Fig. 21 shows the correlation between output power and data TQ. It can be seen from the figure that communication performance is acceptable (Variation less than 10%) when the system works within 400 W to 700 W and $\Delta\tau$ lying between $3\mu s$ and $5\mu s$.

Table 3: Parameters of the prototype IPT system [1]

Operating frequency		22 kHz
Windings type		Spiral
Power MOSFET	V_{DS}	500 V
	V_{GS}	±20 V
Rated output power		700 W
Rated load		40 Ω
Wire diameter		2.0 mm
Grid side inductance		70 uH
Grid side capacitance		0.35 uF
Grid side induction coil outer diameter		165 mm
Grid side winding number		20 turns
Mutual inductance		26.1 uH
Vehicle side inductance		69.4 uH
Vehicle side capacitance		0.35 uH
Vehicle side induction coil diameter		165 mm
Vehicle side winding number		20 turns

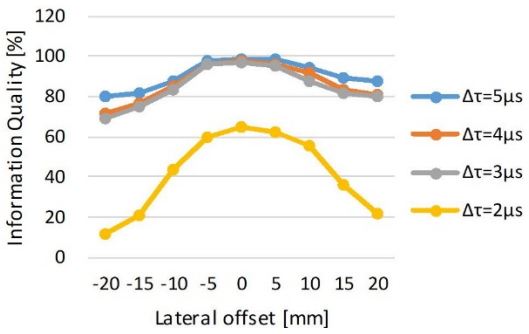


Figure 20. Data TQ for different lateral offsets between two induction pad centers of L_P and L_S with the nominal air gap 25 mm

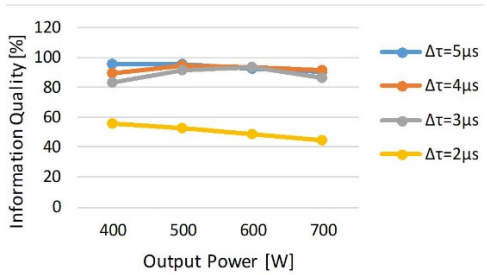


Figure 21. Data TQ vs. output power

Fig. 22 shows that the results of data TQ vs. air gap when the conduction windings L_S and L_P lies between 15 mm and 25 mm. It is quite clear that performance of data transmission drops quickly when the air gap is greater than 25 mm. However, the experimental results were produced by the prototype sysytem. In general, the optimal air gap of data transmission between two windings may change with charcteristics of the system.

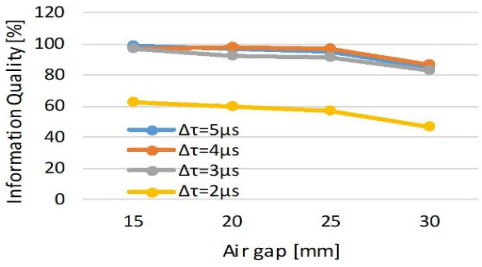


Figure 22. Data TQ vs. air gaps between two induction pads

5. Discussions

While experiments presented in this report demonstrate satisfactory results in either power transmission or data transmission, the power transfer efficiency might drop rapidly due to visible misalignment of two induction pads between vehicle chassis and power station as revealed in Fig. 20. Practically, it is impossible for the driver to move car perfectly so that there is no misalignment between two induction pads. There is a need to develop an adaptive bidirectional wireless power and data transmission control strategy for the optimal TQ under the situation of lateral misalignment between two induction pads. In addition, a clear relationship between data transmission efficiency and physical parameters of the LC tank remains to be investigated.

6. Concluding Remarks

This research report presents a newly developed power and bidirectional data transmission design. The major contribution is that the state of charge of the vehicle battery and vehicle ID information could be mutually communicated through two contactless units. The design exhibits several advantages. For example, the primary (Grid) side of the system can control the output current and achieve steady-state compensation of the secondary (Vehicle) side of the system by wireless transmission. In addition, it costs relatively less while compared with blue tooth or other RF communication technologies because no extra data communication device devices needed. Most importantly, it is a kind of transmission via magnetic field, therefore, there is no serious issue on EMC/EMI. That is, the system is robust against RF interference which makes it to be applicable in the indoor parking lot. To verify applicability, experiments have been conducted with different lateral offsets and vertical shifts between two induction pads which demonstrates performance robustness in power and data transmission.

Acknowledgment

This work was also supported in part by the Innovation and Development Center of Sustainable Agriculture (IDCSA) under the Higher Education Sprout Project, Ministry of Education (MOE) in Taiwan.

References

[1] J. Dai, D. C. Ludois, "A survey of wireless power transfer and a critical comparison of inductive and capacitive coupling for small gap applications" IEEE Trans. Power Electron., 30(11), 6017-6029, 2015. https://doi.org/10.1109/TPEL.2015.2415253

[2] C. C. Huang, H. H. Jian, J. J. Kao, C. L. Lin, Y. T. Kuo, "New Design of Wireless Power and Bidirectional Data Transmission" in 2017 12th IEEE Conference on Industrial Electronics and Applications, Siem Reap, 2017. https://doi.org/10.1109/ICIEA.2017.8282873

[3] X. Qu, H. Han, S. C. Wong, C. K. Tse, W. Chen, "Hybrid IPT topologies with constant current or constant voltage output for battery charging applications" IEEE Trans. Power Electron., 30(11), pp. 6329-6337, 2015. https://doi.org/10.1109/TPEL.2015.2396471

[4] E. A. Mehdi, K. Chma, J. Stefani, "Rapid-charge electric-vehicle stations" IEEE Trans. Power Delivery, 25(3), pp. 1883-1887, 2010. https://doi.org/10.1109/TPWRD.2010.2047874

[5] C. H. Ou, H. Liang, W. Zhuang, "Investigating wireless charging and mobility of electric vehicles on electricity market" IEEE Trans. Ind. Electron., 62(5), pp. 3123-3133, 2015. https://doi.org/10.1109/TIE.2014.2376913

[6] T. Diekhans, R. W. D. Doncker, "A dual-side controlled inductive power transfer system optimized for large coupling factor variations and partial load" IEEE Trans. Power Electron., 30(11), pp. 6320-6328, 2015. https://doi.org/10.1109/TPEL.2015.2393912

[7] U. K. Madawala, M. Neath, D. J. Thrimawithana, "A power–frequency controller for bidirectional inductive power transfer systems" IEEE Trans. Ind. Electron., 60(1), pp. 310-317, 2013. https://doi.org/10.1109/TIE.2011.2174537

[8] D. J. Thrimawithana, U. K. Madawala, M. Neath, "A synchronization technique for bidirectional IPT systems" IEEE Trans. Ind. Electron., 60(1), pp. 301-309, 2013. https://doi.org/10.1109/TIE.2011.2174536

[9] Y. Ma, T. Houghton, A. Cruden, D. Infield, "Modeling the benefits of vehicle-to-grid technology to a power system" IEEE Trans. Power Syst. 27(2), pp. 1012-1020, 2012. https://doi.org/10.1109/TPWRS.2011.2178043

[10] Q. Xu, D. Hu, B. Duan, J. He, "A fully implantable stimulator with wireless power and data transmission for experimental investigation of epidural spinal cord stimulation" IEEE Trans. Neural and Rehabilitation Eng., 23(4) pp. 683-692, 2015. https://doi.org/10.1109/TNSRE.2015.2396574

[11] Y. Jang, M. M. Jovanovic, "A contactless electrical energy transmission system for portable-telephone battery chargers" IEEE Trans. Ind. Electron., 50(3), pp. 520-527, 2003. https://doi.org/10.1109/TIE.2003.812472

[12] S. Jeong, Y. J. Jang, D. Kum, "Economic analysis of the dynamic charging electric vehicle," IEEE Trans. Power Electron., 30(11), pp. 6368-6377, 2015. https://doi.org/10.1109/TPEL.2015.2424712

[13] G. R. Nagendra, L. Chen, G. A. Covic, J. T. Boys, "Detection of EVs on IPT Highways" IEEE Emerging and Selected Topics in Power Electron., 2(3), pp. 584-597, 2014. https://doi.org/10.1109/JESTPE.2014.2308307

[14] T. Bieler, M. Perrottet, V. Nguyen, Y. Perriard, "Contactless power and information transmission" IEEE Trans. Ind. Electron., 38(5), pp. 1266-1272, 2002. https://doi.org/10.1109/TIA.2002.803017

[15] Z. H. Wang, Y. P. Li, Y. Sun, C. S. Tang, X. Lv, "Load detection model of voltage-fed inductive power transfer system" IEEE Trans. Power Electron., 28(11), pp. 5233-5243, 2013. https://doi.org/10.1109/TPEL.2013.2243756

[16] C. S. Wang, O. H. Stielau, G. A. Covic, "Design considerations for a contactless electric vehicle battery charger" IEEE Trans. Ind. Electron., 52(5), pp. 1308-1314, 2005. https://doi.org/10.1109/TIE.2005.855672

[17] C. S. Wang, G. A. Covic, O. H. Stielau, "Power transfer capability and bifurcation phenomena of loosely coupled inductive power transfer system" IEEE Trans. Ind. Electron., 51(1), pp. 148-157, 2004. https://doi.org/10.1109/TIE.2003.822038

[18] Y. Liu, J. J. Yan, H. T. Dabag, P. M. Asbeck, "Novel technique for wideband digital predistortion of power amplifiers with an undersampling ADC" IEEE Trans. Microw. Theory Techn., 62(11), pp. 2604-2617, 2014. https://doi.org/10.1109/TMTT.2014.2360398

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 233-242 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Computational Techniques to Recover Missing Gene Expression Data

Negin Fraidouni*, Gergely Zaruba

Electrical Engineering & Computer Science Department, Wichita State University, Wichita, KS. 67260, USA

ARTICLE INFO

Article history:
Received: 27 June, 2018
Accepted: 29 October, 2018
Online: 10 November, 2018

Keywords:
Gene Expression Prediction
Pearson Correlation Coefficient
Cosine Similarity
Robust Principal Component Analysis
Alternating Direction Method of Multiplier

ABSTRACT

Almost every cells in human's body contain the same number of genes so what makes them different is which genes are expressed at any time. Measuring gene expression can be done by measuring the amount of mRNA molecules. However, it is a very expensive and time consuming task. Using computational methods can help biologists to perform gene expression measurements more efficiently by providing prediction techniques based on partial measurements. In this paper we describe how we can recover a gene expression dataset by employing Euclidean distance, Pearson correlation coefficient, Cosine similarity and Robust PCA. To do this, we can assume that the gene expression data is a matrix that has missing values. In that case the rows of the matrix are different genes and columns are different subjects. In order to find missing values, we assume that the data matrix is low rank. We then used different correlation metrics to find similar genes. In another approach, we employed RPCA method to differentiate the underlying low rank matrix from the sparse noise. We used existing implementations of state-of-the-art algorithms to compare their accuracy. We describe that RPCA approach outperforms the other approaches with reaching improvement factors beyond 4.8 in mean squared error.

1. Introduction

This paper is an extension of works originally presented in ICCABS 2017 (International Conference on Computational Advances in Bio and Medical Sciences) [1] and CSCI 2017 (International Conference on Computational Science and Computational Intelligence) [2].

Almost every cell in an organism's body contain the same genetic information and series of genes, what makes cells different is which genes are expressed at any time. Gene expression is what makes a blood cell different from a liver cell and a normal healthy cell from an abnormal one (like a cancer cell) [3]. Gene expression process has two steps, transcription and translation. Transcription means a particular part of DNA is encoded into messenger RNA (mRNA) and in translation, mRNA is decoded to build a protein that contains a specific series of amino acids. We can measure the gene expression level by measuring the amount of mRNA inside the cell. Each step of the process is regulated by control points that determine the presence and the amount of proteins in any specific cell [4]. Usually a group of genes work accordingly to manage every simple or complex process that control the structure and actions of the cells. This means that group of genes must work together in order to control structure and actions of cells [5]. Knowing this, we can conclude that the gene expression levels should be highly correlated so if the data has missing values, we might be able to predict them based on the correlation between genes.

Recently scientists have the opportunity to find the association between genes and diseases using some methods. Examples of this methods are RNA sequencing, northern blotting, western blotting, DNA microarray, fluorescent in situ hybridization and reporter gene. However the costs of measuring gene expression levels are extremely high and also the complete process needs a huge amount of time [6] which makes it difficult to measure and access this information. Also gene expression data usually suffers from missing values. This can happen due to some reasons like failures in hybridization, noise in data and also data corruption. Missing values in gene expression data can negatively affect gene disease studies [7]. Since due to the huge amount of time and money needed for repeating the measurements, an alternative way is to use computational methods which can be employed to predict the missing values and recover the dataset. So there is a high demand for novel techniques to find the missing values.

The first step to make our model is to store gene expression data in matrices where each row is a different gene, each column

*Negin Fraidouni, Department of Computer Science, The University of Texas at Arlington, Arlington, TX. 76010, USA. Email: neginfraidouni@gmail.com

is a different disease and the entries of the matrix are the gene expression values (mRNA measurements) of corresponding rows and columns [8]. Based on this model we can assume that people with similar diseases show similar expression patterns so the gene expression data matrix must be highly overdetermined and significantly low rank [9].

1.1. Recommendation system

Recommendation systems rely on information filtering in order to deal with data overload by filtering necessary information which is significantly less than total data of users' preferences, interests or observations about an item. Recommender systems have the ability to recommend a new movie to a specific user based on his/her previous preferences [10]. Recently different designs for recommendation systems have been proposed which are based on one of these methods: collaborative filtering method [11], content-based filtering method [12] and hybrid filtering method [13].

1.2. Collaborative Filtering Method

Collaborative filtering method recommends item to users by recognizing users with similar tastes. It combines other ratings in order to recommend new items to each specific user. Collaborative filtering techniques are categorized into two groups: model-based technique and memory-based technique. The main goal for model based method is making a model and extract the necessary part of the data matrix so there is no need to use the entire dataset in order to make predictions [14]. Memory based technique uses previously collected data in order to predict the missing ratings and they use the entire user-item database. The common memory based method is based on nearest neighbors and uses a distance measure metric to find the neighbors [15]. This is also called neighborhood based approach which similar users are grouped together based on their interests [16].

Netflix is an example of recommendation systems that can benefit from collaborative filtering technique. For the Netflix example, the proposed model contains a $m*n$ matrix. Each row of the matrix corresponds to a different user and each column corresponds to a different movie and the entries of the matrix are the ratings users gave to the movies. The data matrix is very sparse because most users usually tend to rate a very small fraction of the movies, the matrix is very sparse. So the goal is to find the hidden pattern and predict the missing values in order to make a recommendation to users for the movies that they have not watched yet.

1.3. Low Rank Matrix Completion

Matrix completion (MC) involves recovering an incomplete matrix where only a small fraction of its entries are known which is significantly smaller than the total size of the matrix. Low rank MC problem can be seen in different practical contexts such as image processing [17], machine learning [18] and bioinformatics [19]. To solve this problem, we should find the lowest rank matrix which is consistent with the known values of the incomplete matrix. We can write:

$$\begin{aligned} &\text{minimize rank (Y)} \\ &\text{such that } R_\omega(Y) = R_\omega(X) \end{aligned} \tag{1}$$

Here X is the incomplete matrix that we want to reconstruct, ω shows the known values such that $(a,b)\in\omega$ if $X_{a,b}$ is known. R_ω is the orthogonal projection matrix where:

$$\left|R_\omega(X)\right|_{a,b} = \begin{cases} X, & (a,b)\in\omega \\ 0, & (a,b)\notin\omega \end{cases} \tag{2}$$

Because the rank minimization problem is NP-hard, this problem can be remodeled as minimizing trace norm or nuclear norm. Nuclear norm is the sum of singular values of the given matrix [20]. The reason is that a rank r matrix with has exactly r singular values which are greater than zero. The nuclear norm of matrix Z is defined as:

$$\|Z\|_* = \sum_{a=1}^{r} \sigma_a \tag{3}$$

Where:

σ_a is the ath singular value (nonzero) of matrix Z and r is the rank of matrix Z.

So we can rewrite the problem as:

$$\begin{aligned} &\text{minimize } (\|Y\|_*) \\ &\text{such that } R_\omega(Y) = R_\omega(X) \end{aligned} \tag{4}$$

The advantage of using nuclear norm over minimizing the rank is that its optimum point can be calculated efficiently and it is convex.

1.4. Robust PCA method (RPCA)

One problem with gene expression datasets is the presence of noise in expression measurements. This happens because of some reasons like different degrees of uniformity, small spots, process errors and also inconsistency in hybridization.

For the aim of showing the most variability of the data for a noise free dataset, we can easily perform PCA using SVD (singular value decomposition). In the presence of noise we can use RPCA in order to reconstruct a low rank matrix and find the sparse noise. Assume that our data matrix E is decomposed as:

$$E = Y + S \tag{5}$$

Where Y is the underlying low rank matrix and S is a sparse matrix capturing noise. Because the number of unknowns to infer for Y and S is considerably higher than known values in E, this problem is overdetermined. So we need to use tractable convex optimization as denoted by:

$$\begin{aligned} &\text{minimize } \|Y\|_* + \lambda\|S\|_1 \\ &\text{subject to } Y + S = E \end{aligned} \tag{6}$$

Where $\|S\|_1 = \sum_{i,j}|S_{i,j}|$ is the ℓ_1 -norm of S and λ is a parameter. This should work even in the situations when the rank of Y is not low rank (when rank is equal to the dimension of the matrix). For RPCA method to work efficiently, we need to know the location of the non-zero entries in matrix S. Problem (6) can

be solved at a cost not so much higher than classical PCA [21]. One of the methods that can be employed here is the Alternating Direction Method of Multipliers (ADMM) that we will summarize in the next section.

1.5. The ADMM method

The ADMM method is a powerful method because it mixes two methods of multipliers and dual ascent. The algorithm solves the problems in the form:

$$\begin{aligned} &\min \ \ f(a)+g(b) \\ &\text{such that } Xa+Yb=z \end{aligned} \tag{7}$$

Where f and g are both convex. The optimal value for the problem above is defined as:

$$p^{*}=\inf\{f(a)+g(b)\mid Xa+Yb=z\} \tag{8}$$

Here the augmented Lagrangian is:

$$\begin{aligned} &L_{\rho}(a,b,m)=f(a)+g(b)+m^{T}(Xa+Yb-z)+ \\ &\frac{\rho}{2}\ \|Xa+Yb-z\ \|_{F}^{2} \end{aligned} \tag{9}$$

Where m is the Lagrangian multiplier and $\rho>0$ is a parameter. ADMM method consists of the multiple iterations as denoted below:

$$\begin{aligned} a^{k+1} &= arg_{a} minL_{\rho}(a^{k},b^{k},m^{k}) \\ b^{k+1} &= arg_{b} minL_{\rho}(a^{k+1},b^{k},m^{k}) \\ m^{k+1} &= m^{k}+\rho(Xa^{k+1}+Yb^{k+1}-z) \end{aligned} \tag{10}$$

The algorithm consists of multiple steps:

1. An a-minimizing step
2. A b-minimizing step and
3. A variable update.

In the last step (variable update step) the step size is equal to the m (the augmented Lagrangian parameter).

2. Methods

2.1. Correlation based Matrix Completion method

The main goal of the correlation based matrix completion method (CMC) is finding correlation between genes (neighbors). To predict a value for a missing entries, we need to consider all other subjects' expression values. If a gene is more similar to the one with a missing value, its expression value has more impact on the predicted value. For finding correlation between genes, we will use Pearson correlation coefficient (PCC), Euclidean distance (ED) and Cosine similarity (CS).

Pearson Correlation Coefficient

PCC is a common measure of linear dependency between two variables. The PCC can take any value from 1- (means negative association) to 1 (means positive association) with 0 indicating orthogonality. The PCC can be calculated by:

$$PCC=\frac{\sum_{k=1}^{m}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\sum_{k=1}^{m}(x_k-\overline{x})^2}\sqrt{\sum_{k=1}^{m}(y_k-\overline{y})^2}} \tag{11}$$

Or:

$$PCC=\frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2-(\Sigma x)^2][n\Sigma y^2-(\Sigma y)^2]}} \tag{12}$$

Where X and Y are two datasets.

Euclidean Distance

Euclidean distance (ED) is a metric that if x and y have zero distance, then x = y holds. ED between two points (x and y) can be calculated by:

$$ED=|x-y|=\sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} \tag{13}$$

Cosine Similarity

The Cosine Similarity (CS) is a metric of the cosine of the angle between two vectors. This is a measurement of orientation instead of magnitude. Like PCC, CS can take any value from 1- (means negative association) to 1 (means positive association) with 0 indicating orthogonality. The CS value can be calculated as shown below:

$$cos\theta=\frac{\vec{x}.\vec{y}}{\|\vec{x}\ \|.\ \|\vec{y}\ \|}=\frac{\sum_{k=1}^{m}x_k y_k}{\sqrt{\sum_{k=1}^{m}x_k^2}\sqrt{\sum_{k=1}^{m}y_k^2}} \tag{14}$$

And the dot product:

$$\vec{x}.\vec{y}=\|\vec{x}\ \|.\ \|\vec{y}\ \|.cos\theta \tag{15}$$

CMC Approach

Here we explain how CMC approach works in order to find missing values in partially known matrices. Let us assume that we have a complete (means all the entries are known) low rank matrix Y. Now if we randomly remove some of the values, the problem becomes how to find missing values in a way that there was not a huge difference between the original values and the predictions. We will use PCC, ED and CS in order to reconstruct matrix X and the reconstructed matrices when using each of the aforementioned similarity metrics are Y_P, Y_E and Y_C respectively.

This is how the CMC method works:

- First a PCC, an ED and a CS value should be calculated for each pair of the genes.
- When there is a missing value, we need to calculate the mean of all known entries in that column weighted by how correlated they are (based on their PCC, ED and CS values).
- Finally we will measure the accuracy of each method by calculating the error between the reconstructed matrices and the original matrix.

For finding a missing value at *Y (m,n)* we will calculate:

$$Y_P(m,n) = \frac{\sum_{r=1}^{K} PCC'(m,n,r) * y''_{r,n}}{\sum_{r=1}^{N} x(r,n)} \quad (16)$$

$$Y_E(\mathrm{m,n}) = \frac{\sum_{r=1}^{K} ED'(m,n,r) * y''_{r,n}}{\sum_{r=1}^{K} x(r,n)} \quad (17)$$

$$Y_C(m,n) = \frac{\sum_{r=1}^{K} CS'(m,n,r) * y''_{r,n}}{\sum_{r=1}^{K} x(r,n)} \quad (18)$$

Where *r! = m* is matrix *Y*'s row, and:

$$PCC'_{r,n} = \begin{cases} PCC(m,r) & (r,n) \in \omega \\ 0 & (r,n) \notin \omega \end{cases} \quad (19)$$

$$ED'_{r,n} = \begin{cases} ED(m,r) & (r,n) \in \omega \\ 0 & (r,n) \notin \omega \end{cases} \quad (20)$$

$$\mathrm{CS}'_{r,n} = \begin{cases} \mathrm{CS}(m,r) & (r,n) \in \omega \\ 0 & (r,n) \notin \omega \end{cases} \quad (21)$$

$$y''_{r,n} = \begin{cases} y'_{r,n} & (r,n) \in \omega \\ 0 & (r,n) \notin \omega \end{cases} \quad (22)$$

$$x(r,n) = \begin{cases} 1 & (r,n) \in \omega \\ 0 & (r,n) \notin \omega \end{cases} \quad (23)$$

Algorithm 1 shows the pseudo code for correlation based matrix completion approach.

Algorithm 1. Correlation based matrix completion approach

```
Input: Y, ω
    For row m of Y:
        For row n of Y:
            If m != n :
                Find PCC (m,n)
                Find ED(m,n)
                Find CS (m,n)

    For row m of Y:
        For row r of Y:
            If (m,r) in ω:
                Y_p(m,r) = Y(m,r)
                Y_e(m,r) = Y(m,r)
                Y_c(m,r) = Y(m,r)
            Else if (m,r) not in ω:
                Y_p(m,r) = 0
                Y_e(m,r) = 0
                Y_c(m,r) = 0
                x = 0
                for row p of Y:
                    if (p,n) in ω:
                        Y_p(m,r) += Y_p(r,n) *PCC(m,p)
                        Y_e(m,r) += Y_e(r,n) * ED(m,p)
                        Y_c(m,r) += Y_c(r,n) * CS(m,p)
                        x++
                Y_p(m,r) /= x
                Y_e(m,r) /= x
                Y_c(m,r) /= x
output: Y_P, Y_E, Y_C
```

2.2. Convex Optimization Formulation of RPCA method

The ADMM formulation of our RPCA model (6) is defined by:

$$L_\rho(Y,S,m) = \|Y\|_* + \lambda \|R_\omega(S)\|_1 + m^T(R_\omega(E-Y-S)) + \frac{\rho}{2}\|R_\omega(E-Y-S)\|_F^2 \quad (24)$$

In each iteration we repeat:

Updating matrix Y

We update Y by:

$$Y^k = min_Y \|Y\|_* + \frac{\rho}{2}\|R_\omega(E-S^{k-1}-Y^{k-1})\|_F^2 + m^T(R_\omega(E-S^{k-1}-Y^{k-1})) \quad (25)$$

Which can be rewritten as:

$$min_Y \|Y\|_* + \frac{\rho}{2}\|P_\Omega(Y^k+S^k-E)-\frac{m}{\rho}\|_F^2 \quad (26)$$

For solving above problem, we can use a soft thresholding operation from [22]. So the problem would become:

$$Y^k = shrink(A^{k-1}, \rho^{-1})$$
$$A^{k-1} = (E-S^{k-1}+\frac{m}{\rho^{k-1}}) \quad (27)$$

Where ρ^{-1} is the step size which decreases the singular values of matrix A and the shrink is a soft-thresholding operator and can be defined as:

$$shrink(M,\tau) := \sum_{k=1}^{r} u_i, max(\sigma_k - \tau, 0) v_k^T$$
$$M = \sum_{k=1}^{r} (u_k \sigma_k v_k^T) \quad (28)$$

Where σ_k is the singular values and u_k is the left and v_k is the right singular vectors of matrix M.

Updating matrix S:

After updating Y, we can update S through:

$$S^k = min_S \lambda \; \| R_\omega(S^{k-1}) \|_1 + m^T (R_\omega(E - S^{k-1} - Y^k)) + \frac{\rho}{2} \| R_\omega(E - S^{k-1} - Y^k) \|_F^2 \quad (29)$$

Which can be rewritten as:

$$min_S \lambda \; \| R_\omega(S^{k-1}) \|_1 + \frac{\rho}{2} \| R_\omega(Y^k - E + S^{k-1}) - \frac{m}{\rho} \|_F^2 \quad (30)$$

To solve the above problem, we can use a shrinkage operator:

$$\begin{cases} S_{ij} = H_{\frac{\lambda}{\rho}}(E - Y^k + \frac{m}{\rho}) & (i,j) \in \omega \\ S_{ij} = 0, & (i,j) \notin \omega \end{cases} \quad (31)$$

Where $H_{\frac{\lambda}{\rho}}$ is the shrinkage operator discussed in [23] and can be calculated by:

$$H_\sigma(S_{ij}) = \begin{cases} S_{ij} - \sigma, & S_{ij} > \sigma \\ S_{ij} + \sigma, & S_{ij} < -\sigma \\ 0 & \text{Otherwise} \end{cases} \quad (32)$$

We can assume that entries in matrix S that represent missing values are equal to zero.

Updating m:

After updating Y and S, we can update m by:

$$m^k = m^{k-1} + \rho(E - Y^k - S^k) \quad (33)$$

Algorithm 2 shows the pseudo code for solving RPCA problem using ADMM.

Algorithm 2. Solving RPCA problem using ADMM

Input: E, ρ, λ, ε
While $\|E - Y^k - S^k\|_F > \varepsilon$:

Updating matrix Y:
$Y^k = \arg_Y \min L\rho(Y^{k-1}, S^{k-1}, m^{k-1})$
$Y^k = shrink\ ((E - S^{k-1} + \frac{m}{\rho}), \rho^{-1})$
$(U,S,V) = SVD\ (E - S^{k-1} + \frac{m}{\rho})$
For singular values σ in S:
If $\sigma < \frac{1}{\rho}$:
$\sigma = 0$
$Y^k = U\ S\ V^T$

Updating matrix S:
$S^{k+1} = \arg_S \min L\rho(Y^{k+1}, S^k, m^k)$
for row p of S:
for column r of S:
if (p,r) in ω:
$S_{pr} = H\frac{\lambda}{\rho}\ (E - Y^k + \frac{m}{\rho})$
else:
$S_{pr} = 0$

Updating m:
$m^k = m^{k-1} + \rho(E - S^k - Y^k)$

output: Y^k and S^k

3. Competitive methods

3.1. K-nearest Neighbors method

K-nearest neighbors (KNN) is one of the most essential classification algorithms in machine learning. It can be widely used in real-life scenarios since it does not make any assumption about the distribution of the data. The model representation for KNN is the entire dataset and it can make predictions using the training data set directly. When there is a missing value, prediction can be made by searching through the dataset for the K most similar neighbors and the result is the weighted average of those neighbors [24]. To determine which K neighbors are the most similar ones, a distance measure should be used and Euclidean distance is the most popular one for real-valued variables.

3.2. Nuclear Norm Minimization

There are various numerical methods available to solve(4). The important problem is that because of the high dimensionality aspect of biological data, many numerical methods fail to solve the problem efficiently. Kapur et al. [25] used a method called soft thresholding operator which can scale well on large datasets. So the problem would become:

$$\min \; \tau \|Y\|_* + \frac{1}{2} \|Y\|_F \quad \text{such that } R_\omega(Y) = R_\omega(X) \quad (34)$$

Where $\|Z\|_F = \sqrt{\sum_{i=1}^{m} \sum_{i=1}^{n} |z_{i,j}|^2}$ is the Frobenius norm and τ is the thresholding parameter and it should be greater than 0. We can reconstruct the expression matrix iteratively so the kth iteration would be:

$$Y^k = \text{shrink}(M^{k-1}, \tau)$$
$$M^k = M^{k-1} + \delta_k R_\omega(X - Y^k) \quad (35)$$

Where shrink is the soft thresholding operator [22]. The parameter δ_k is the step size and the parameter τ minimizes the rank by decreasing the singular values. The shrink operator is defined by:

$$\text{shrink}\ (M,\tau) \coloneqq \sum_{i=1}^{r} \max(\sigma_i - \tau, 0)\ u_i v_i^T$$
$$M = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{36}$$

Where u_i is the left and v_i is the right singular vectors of data matrix M. In each iteration the SVD of matrix M is calculated and those singular values that are smaller than τ parameter, will be set to zero. The new matrix M will be reconstructed. Algorithm 3 shows the pseudo code for this method.

Algorithm 3. Nuclear Norm Minimization Problem

```
Input: Y, ω, ε
δ = 1.2 *(mn)/|ω|
τ = 5 *(mn)^0.5
Shrink(Y, τ)
    (U,S,V) = SVD (Y)
    For singular values σ in S:
        If σ < τ:
            σ = 0
    M = U S V^T

Minimize(Y, ω)
    For row a of Y:
        For column b of Y:
            If (a,b) in ω:
                Rω (Y) = Y
            Else:
                Rω (Y) = 0
    M^0 = 0
    k = 1
    while ‖Rω (Y^k - X)‖_F / ‖Rω (X)‖_F < ε:
        Y^k = shrink (Y^(k-1), τ)
        M^k = M^(k-1) + Rω (X - Y^k)
        k++
output: X^k
```

3.3. Singular Value Thresholding Algorithm (SVT)

This approach considers using a Robust PCA approach in order to reconstruct a low rank matrix from noisy measurements.

$$min_{A,E} \lambda \|E\|_1 + \|A\|_* + 2\tau^{-1} \|E + A\|_F^2$$
$$\text{such that: } D = E + A \tag{37}$$

Where:

- D is the noisy dataset.
- A is the low rank matrix.
- E is the noise and it assumed that it only affect a fraction of the data (E is sparse).
- τ is a scalar and $\tau > 0$.

We can apply the Lagrangian multiplier Y in order to replace the equality constraint:

$$L(\mathrm{E},\mathrm{A},\mathrm{Y}) = \lambda\|E\|_1 + \|A\|_* + 2\tau^{-1}\|E + A\|_F^2 + \frac{1}{\tau}\langle Y, D - A - E\rangle \tag{38}$$

Then in each iteration A, E and Y will be updated by minimizing (38) with respect to A, E and Y. Algorithm 4. Shows the pseudo code for SVT approach where σ is the step size.

Algorithm 4. The SVT Algorithm

```
Input: τ,D,λ
While not converged:
    (U,S,V) = SVD(Y^k)
    A^k= arg min_x τ‖D‖_* + 1/2 ‖D-E^(k-1)‖_F
    E^k= arg min_x τ‖D‖_1 + 1/2 ‖D-A^k‖_F
    Y^k= Y^(k-1) + σ_k (D - A^k - E^k)
End while
output: A = A^k, E = E^k
```

3.4. Exact Augmented Lagrangian Multiplier (ELAM)

ELAM method was proposed in [26] and can be used for solving Robust PCA problem. To solve the problem can apply the Lagrangian multiplier as denoted below:

$$X = (\mathrm{A},\mathrm{E})$$
$$f(X) = \|A\|_* + \lambda\|E\|_1 \tag{39}$$
$$h(X) = D - A - E$$

And the Lagrangian function is:

$$L(A,E,Y) = \|A\|_* + \lambda\|E\|_1 + \frac{\mu}{2}\|D - A - E\|_F^2 + Y^T(D - A - E) \tag{40}$$

Algorithm 5. Shows the pseudo code for ELAM method.

Algorithm 5. Exact Augmented Lagrangian Multiplier (ELAM)

```
Input: matrix D, λ
While not converged:
    (A^(k+1), E^(k+1)) = arg min_(A,E) L(A,E,Y)
        While not converged:
            (U,S,V) = SVD(D - E^k + Y/μ)
            A^(k+1)= arg min_x 1/μ‖D‖_* + 1/2 ‖D-E^k‖_F
            E^(k+1)= arg min_x λ/μ‖D‖_1 + 1/2 ‖D-A^(k+1)+λ/μ‖_F
        End while
    Y^(k+1)= Y^k + μ^k(D - A^(k+1) - E^(k+1))
    k++
End while
output: A = A^(k+1), E = E^(k+1)
```

3.5. Inexact Augmented Lagrangian Multiplier (ILAM)

ILAM method was proposed in a study by Lin et al [27]. The RPCA problem is closely connected to MC problem so the MC can be formulated as:

$$\begin{aligned} &\text{minimize}_A \|A\|_* \\ &\text{such that } E + A = D,\ R_\omega(E) = 0 \end{aligned} \tag{41}$$

$R_\omega(E) = 0$ means that E is zero at indices where the value is known and the augmented Lagrangian is:

$$\begin{aligned} L(E,A,Y) &= \|A\|_* + \frac{\mu}{2}\|D - E - A\|_F^2 \\ &+ Y^T(D - E - A) \end{aligned} \tag{42}$$

So the ILAM approach can be used for MC problem. The pseudo code for ILAM method is described below.

```
Algorithm 6. Exact Augmented Lagrangian Multiplier
(ELAM)
```

```
Input: matrix D, ω, λ
While not converged:
    (A^k, E^k) = arg min_{A,E} L(E,A,Y)
    (U,S,V) = SVD(D − E^k + Y/μ)
    A^k= arg min_x (1/μ)‖D‖* + (1/2) ‖D−E^{k−1}‖F
    E^k= arg min_x (λ/μ)‖D‖1 + (1/2) ‖D−A^k+λ/μ‖F
    Rω (E) = 0
    Y^k= μ^k(D − E^k− A^k)+ Y^{k−1}
    k++
End while
output: E = E^k, A = A^k
```

4. Evaluation

Here we measure the accuracy of different approaches as they apply to biomedical data MC. All of the following experiments were performed using Python 3.6 and Matlab 2016 on an Intel Core i7 PC running Windows 10 with 16GB main memory.

4.1. Datasets

We downloaded and used 4 gene expression datasets from NCBI (National Center for Biotechnology Information) for our experiments. We used the following gene expression datasets:

Lung Cancer Study

Title of this study is: "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival". Scientists found that smoking tobacco is the reason for the most of lung cancer cases, but the exact details of this process is still unknown. In this study Landi et al. used 135 tissue samples of adenocarcinoma and non-involved lung tissue from 3 groups (current, former and never smokers). They found out that expression of some genes is significantly different in smokers and non-smokers. The lung cancer dataset has 22283 rows- and 107 columns [28].

Dementia Study

Title of this study is: "Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration". The symptoms of frontotemporal lobar degeneration is progressive decline in language and function. Despite the excessive research on the reason for this disease, its mechanisms remain unknown. Plotkin et al. isolated postmortem brain samples from normal controls, patients with mutations in progranulin gene and patients without mutations in progranulin gene. The dementia dataset has 22277 rows and 56 columns [29].

Autism Study

Title of this study is: "Autism and increased paternal age related changes in global levels of gene expression regulation". Autism is a neurodevelopmental disorder and it is the results of transcription factor mutations that can change the gene expression regulation. In this study Alter et al. analyzed gene expression values of 82 subjects with autism and 64 controls. The results showed that autism and increased paternal age can change the gene expression regulation. The Autism dataset has 54613 rows and 146 columns [30].

Bladder Cancer Study

Title of this study is: "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer". In this study Riester et al. analyzed data of patients with bladder cancer (n = 93) and measured the gene expression. The bladder cancer dataset has 54675 rows and 93 columns [31].

4.2. Calculating Error

We used two metrics to determine how accurate the MC algorithm is. So we start with a known matrix Y, remove a random portion of it (i.e., simulating missing entries), and then try to reconstruct the matrix Y'.

Relative Error

Relative error can be used to describe accuracy; specifically, how accurate a measurement is compared to the true value. We use the relative error (RE) which can be calculated as below:

$$\text{Relative Error} = \frac{\|Y - Y'\|_F}{\|Y\|_F} \tag{43}$$

Where Y is the original matrix and Y' is the reconstructed matrix.

Mean Square Error

We will also use mean squared error (MSE) to measure the accuracy of the different approaches. We can find MSE by calculating the mean of the squares of the deviations, which is the

Y and Y' is:

$$\text{Mean Squared Error} = \frac{\|Y - Y'\|_F}{n_1 * n_2} \quad (44)$$

5. Results

We used the complete datasets as starting points for all experiments and removed a random set of values from the data matrices. The resulting incomplete matrices were then used as inputs to the algorithms to predict the unknowns (missing values). In order to measure the accuracy of the different methods, we used two metrics to compare the predicted and the original matrices. We will evaluate the performance of the methods when the degree of missing values changes. To do this, for each dataset we made nine incomplete matrices with 10% to 90% missing values. For each matrix with varied proportion of missing values, we employed the algorithm 10 times so in figures 1, 2 and 3, each data-point shows an average of ten different experiments with randomly removed values from the original matrix.

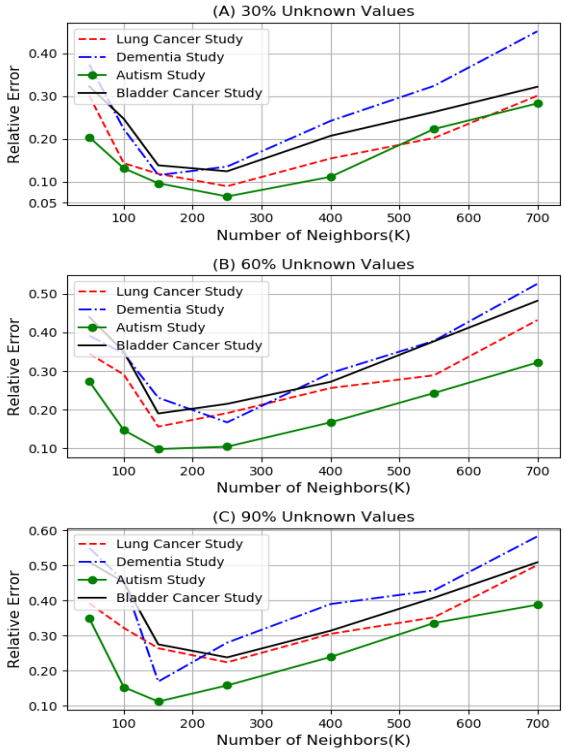


Figure 1. Comparison of relative errors for different values of k for KNN method.

For the aim of finding the best k for KNN method, we varied the value of k from the list 50, 100, 150, 250, 400, 550, 700 and calculated the relative error in cases where 30%, 60% and 90% of the values are unknown. As Figure 1 shows, when the value is around 150-250, the relative error is the least so we selected 200 as the number of neighbors. The results of the comparison between KNN algorithm and PCC- CMC method are summarized in table 1. KNN method is extremely slow when the size of the dataset is large.

Table 1. Comparison of relative error averages for 4 datasets of KNN method and PCC-CMC

	KNN (k = 200)	PCC-CMC
30% Unknowns	0.103	0.047
60% Unknowns	0.169	0.069
90% Unknowns	0.225	0.102

We compared the performance of nuclear norm minimization to PCC, ED and CS based correlation approach on 4 NCBI-GEO datasets (Dementia, Autism, Lung cancer and Bladder cancer) and results are displayed in figure 2. The horizontal axis of all graphs (A - H) represents the ratio of the missing entries. The vertical axis of graphs A, C, E and G represents the relative error (Eq. 43) and the vertical axis in graphs B, D, F and H shows mean squared error (Eq. 44). The performance of PCC-CMC approach is shown by the dotted red line while the black, the dotted blue and the green lines depict the performances of the ED-CMC, CS-CMC and nuclear norm minimization approaches respectively. As the figure 2. Shows, for all four datasets, the PCC-CMC approach consistently beats the nuclear norm minimization approach. The nuclear norm minimization (green line) represents an increasingly growing error but the error of the PCC-CMC approach (red line) shows a decreasing acceleration. The CS-CMC (blue line) also represents improvements when compared to nuclear norm minimization but the ED-CMC approach does not show any improvements. The relative error and mean squared error of PCC-CMC grew very much slower than that of the nuclear norm minimization approach in cases of an increase in the ratio of missing entries. We can explain the different results of PCC-CMC and CS-CMC by this hypothesis that PCC might be better in catching the genes correlation compared to CS.

Based on our results, The PCC-CMC approach has higher accuracy compared to other approaches. In the case of 90% missing values for relative error, in the best case, PCC-CMC outperforms CS-CMC, ED-CMC, and Nuclear norm minimization by a factor of 1.4, 2 and 2.2 respectively. In the worst case PCC-CMC outperforms CS-CMC, ED-CMC, and Nuclear norm minimization by a factor of 1.2, 1.6 and 1.7 respectively. When looking at MSE, PCC-CMC outperforms the other three approaches by as much as a factor of 1.7, 2.3 and 2.4 respectively (for the same order as previously) and in the worst case we get an improvement factor of 1.1, 1.3 and 1.4 respectively.

We then compared the performance of the featured ADMM approach to SVT, ELAM, ILAM and PCC-CMC approach for our 4 datasets and the results are presented in figure 3. The horizontal axis in all graphs (A - H) represents the ratio of the missing entries. The vertical axis in graphs A, C, E and G shows the relative error and the vertical axis in graphs B, D, F and H shows mean squared error. The performance of the featured ADMM approach is shown by the dotted red line while the black, the dotted blue, the green and the gray lines represent the performances of the SVT, ELAM, ILAM and PCC-CMC approaches respectively. As Figure 3. Shows, for all four datasets, the ADMM approach beats the other four approaches. The black,

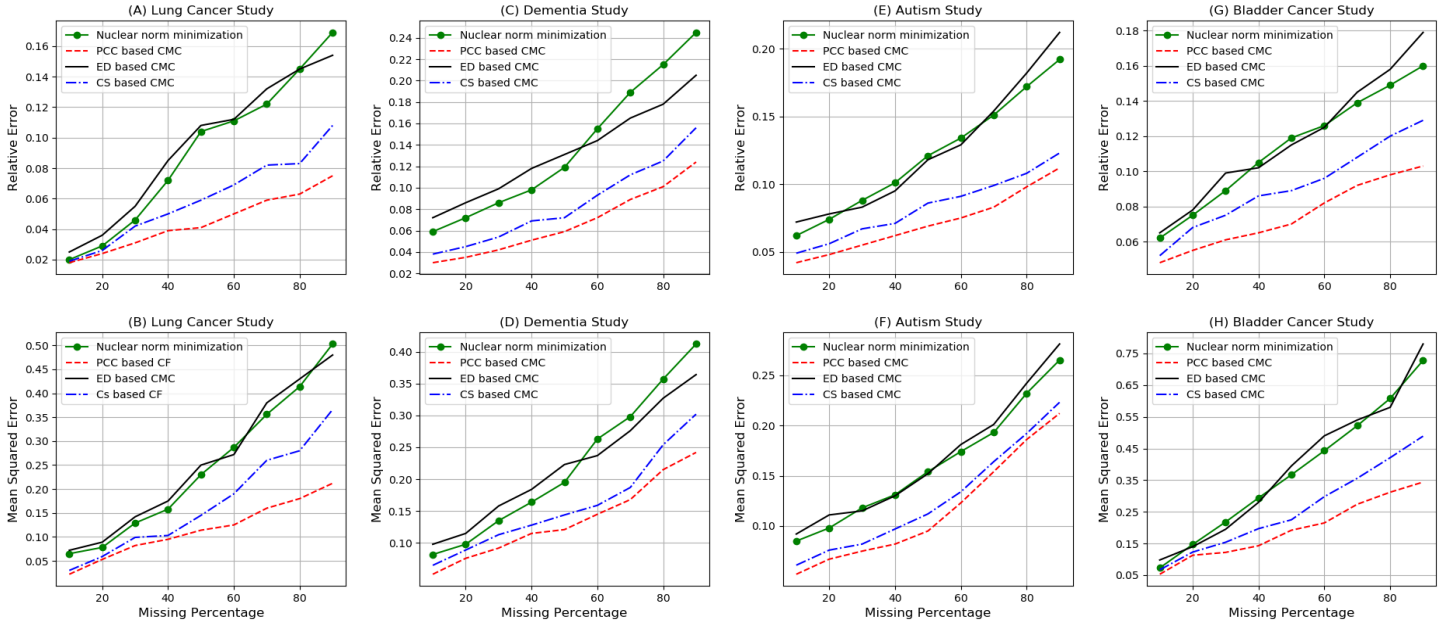


Figure 2. Comparison of the performance of 4 different methods on NCBI-GEO dataset

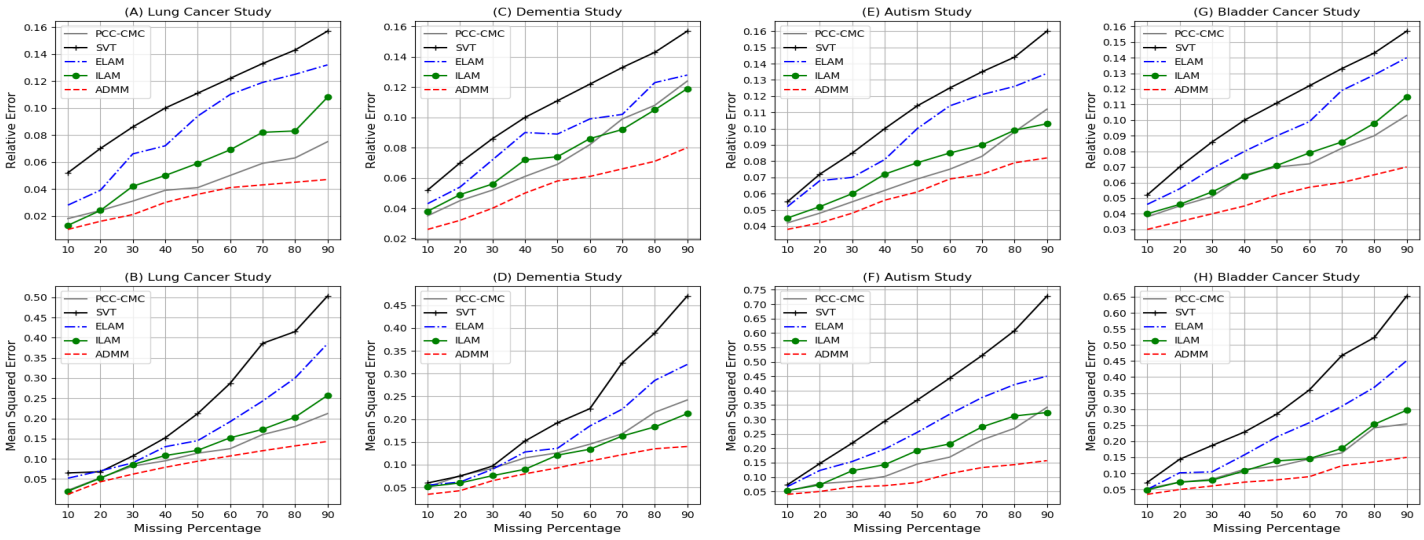


Figure 3. Comparison of 5 different approaches on 4 NCBI-GEO datasets.

blue and green lines show increasingly growing errors but the errors of the red line show a decreasing acceleration. The mean squared error of the ADMM approach grows much slower than that of the other four in cases of an increase in the ratio of missing entries.

Based on our results, The ADMM approach has higher accuracy especially in the cases where the matrix has more missing values. In the case of 90% missing values for relative error, in the best case, ADMM outperforms PCC-CMC, ILAM, ELAM, and SVT by a factor of 1.85, 3, 3.3, and 4 respectively. In the worst case ADMM outperforms PCC-CMC, ILAM, ELAM, and SVT by a factor of 1.4, 1.3, 1.7 and 2 respectively. When looking at MSE, ADMM outperforms the other four approaches by as much as a factor of 2.2, 2.3, 3 and 4.8 respectively (for the same order as previously) and in the worst case we get an improvement factor of 1.6, 1.6, 2.3 and 3.4 respectively.

6. Conclusion

In this paper we employed three similarity metrics (Pearson Correlation Coefficient, Euclidean distance and Cosine Similarity) and Robust Principal Component Analysis (RPCA) on gene expression datasets. In section 2, we described the

correlation based MC approach, the RPCA approach and also we briefly explained the Alternating Direction Method of Multipliers (ADMM) algorithm. We used 4 different gene expression datasets from NCBI and in the first step, we randomly removed a fraction of the entries (from 10% - 90%). So for each dataset we had 9 incomplete matrices that we aim to recover and predict the missing values using one the aforementioned approaches. When we measured the accuracy of the three correlation based approaches, K-nearest neighbors and a recent nuclear-norm minimization based approach, we found out the PCC-CMC approach outperforms the other methods.

In another experiment we evaluate the performance of ADMM approach. To do this, we described three well known algorithms that can be used when recovering low rank matrices and we compared the performances of them. We found that ADMM approach outperforms the other approaches.

This paper can provide an inspiration for developing new approaches especially in gene expression studies and also has implications to recommender systems. There is a high demand for new efficient and fast methods to reduce the huge amount of time and resources that is often needed for gene expression studies. Using such computational methods can help biologists find missing values in partially known gene expression datasets and also can help identify promising directions for studies based on partial measurements in gene expression experiments.

References

[1] N. Fraidouni and G. Zaruba, "A correlation based matrix completion approach to gene expression prediction," in *7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, Orlando, FL, 2017.

[2] N. Fraidouni and G. Zaruba, "A Robust Principal Component Analysis via Alternating Direction Method of Multipliers to Gene-Expression Prediction," in *Proceedings of the 2017 International Conference on Computational Science and Computational Inteligence (CSCI)*, Las Vegas, NV, 2017.

[3] R. Hammamieh, N. Chakraborty, A. Gautam and S. Muhie, "Whole-genome DNA methylation status associated with clinical PTSD measures of OIF/OEF veterans," Transl Psychiatry, PMID: 28696412, 2017.

[4] Y. Bromberg, "Chapter 15: Disease gene prioritization," *PLoS Computational Biology,* 2013.

[5] A. Wong, W. H. Au and K. Chen, "Discovering high-order patterns of gene expression levels," *Journal of Computational Biology,* pp. 625-637, 2008.

[6] S. Welsh and S. Kay, "Reporter gene expression for monitoring gene transfer," *Current Opinions in Biotechnology,* pp. 617-622, 1997.

[7] A. W. Liew, N. Law and H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," *Briefings in Bioinformatics,* vol. 12, no. 5, pp. 495-513, 2011.

[8] V. Gligorijevic and N. Przulj, "Computational Methods for Integration of Biological Data," *Springer International Publishing,* pp. 137-178, 2016.

[9] X. Feng and X. He, "Inference on low rank data matrices with applications to microarray data," *The Annuals of Applied Statistics,* pp. 217-243, 2010.

[10] F. O. Isinkaye, Y. O. Folajimi and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal,* vol. 16, no. 3, pp. 261-273, 2015.

[11] A. M. Acilar and A. Arslan, "A collaborative filtering method based on Artificial Immune Network," *Expert Systems with Applications,* vol. 36, no. 4, pp. 8324-8332, 2009.

[12] L. S. Chen, F. H. Hsu, M. C. Chen and Y. C. Hsu, "Developing recommender systems with the consideration of product profitability for sellers," *International Journal of Geographical Information Science,* vol. 187, no. 4, pp. 1032-1048, 2008.

[13] M. Jalali, N. Mustafa, M. Sulaiman and A. Mamay, "WEBPUM: a web-based recommendation system to predict user future movement," *Expert Systems with Applications,* vol. 37, no. 9, pp. 6201-6212, 2010.

[14] M. Ekstrand, J. T. Reidl and J. Konstan, "Collaborative Filtering Recommender Systems," *Foundations and Trends in Human–Computer Interaction,* vol. 4, pp. 81-173, 2011.

[15] Y. El Madani El Alami, E. H. Nfaoui and O. El Beqqali, "Improving Neighborhood-Based Collaborative Filtering by A Heuristic Approach and An Adjusted Similarity Measure," in *Proceedings of the International Conference on Big Data, Cloud and Applications*, Tetuan, Morocco, 2015.

[16] F. Alqadah, C. Reddy and J. Hu, "Biclustering neighborhood-based collaborative filtering method for top-n recommender systems," *Springer-Verlag London,* 2014.

[17] X. Zhou, C. Yang, H. Zhao and W. Yu, "Low-Rank Modeling and Its Applications in Image Analysis," *ACM Computing Surveys,* vol. 47, no. 2, 2014.

[18] J. Gillard and K. Usevich, "Structured low-rank matrix completion for forecasting in time series analysis," *Elsevier,* 2018.

[19] E. C. Lai, P. Tomancak, R. W. Williams and G. M. Rubin, "Computational identification of Drosophila MicroRNA genes," *Genome Biology,* vol. 4, 2003.

[20] E. Candes and B. Recht, "Exact matrix completion via convex optimization," *Applied and Computational Mathematics,* 2008.

[21] J. Wright, Y. Peng and Y. Ma, "Robust principal component analysis: exact recovery of corrupted low rank matrices by convex optimization," in *NIPS*, 2009.

[22] J. F. Cai, E. J. Candes and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal of Optimization,* vol. 20, no. 4, pp. 1956-1982, 2010.

[23] I. Daubechies, M. Defrise and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Communications on Pure and Applied Mathematics,* 2008.

[24] S. Taneja, C. Gupta, K. Goyal and D. Gureja, "An enhanced K-nearest neighbor algorithm using information gain and clustering," in *Fourth International Conference on Advanced Computing & Communication Technologies*, 2014.

[25] A. Kapur, K. Marwah and G. Alterovitz, "Gene expression prediction using low-rank matrix completion," *BMC Bioinformatics,* pp. 1634-1654, 2016.

[26] Z. Lin, M. Chen and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Mathematical Programming,* 2010.

[27] Z. Lin, M. Chen and Y. Ma, "Linearized Alternating Direction Method with Adaptive Penalty for Low Rank Representation," in *Conference on Neural Information Processing Systems (NIPS)*, 2011.

[28] M. T. Landi, T. Dracheva, M. Rotunno and J. D. Figueroa, "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," PLoS One, PMID: 18297132, 2008.

[29] A. S. Chen-Plotkin, F. Geser, J. B. Plotkin and C. M. Clark, "Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration," *Human Molecular Genetics,* vol. 17, no. 10, p. PMID: 18223198, 2008, PMID: 18223198.

[30] M. D. Alter, R. Kharkar, K. E. Ramsey and D. W. Craig, "Autism and increased paternal age related changes in global levels of gene expression regulation," *Plos One,* vol. 6, no. 2, p. PMID: 21379579, 2011.

[31] M. Riester, J. M. Taylor, A. Feifer and T. Koppie, "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer," Clinical Cancer Research, PMID: 22228636, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 235-247 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Towards an Efficient Federated Cloud Service Selection to Support Workflow Big Data Requirements

Mohamed Adel Serhani[1*], Hadeel Al Kassabi[2,1], Ikbal Taleb[2]

[1]*Information Systems and Security Department, College of Information Technology, 15551, UAE*

[2]*Concordia Institute for Information System Engineering, Concordia University, H3G 1M8, Canada*

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 27 September, 2018
Online: 08 October, 2018

Keywords:
Big Data
Cloud
Service selection
Workflow
QoS

ABSTRACT

Cloud Computing is considered nowadays an attractive solution to serve the Big Data storage, processing, and analytics needs. Given the high complexity of Big Data workflows and their contingent requirements, a single cloud provider might not be able alone to satisfy these needs. A multitude of cloud providers that offer myriad of cloud services and resources can be selected. However, such selection is not straightforward since it has to deal with the scaling of Big Data requirements, and the dynamic cloud resources fluctuation. This work proposes a novel cloud service selection approach which evaluates Big Data requirements, matches them in real time to most suitable cloud services, after which suggests the best matching services satisfying various Big Data processing requests. Our proposed selection scheme is performed throughout three phases: 1) capture Big Data workflow requirements using a Big Data task profile and map these to a set of QoS attributes, and prioritize cloud service providers (CSPs) that best fulfil these requirements, 2) rely on the pool of selected providers by phase 1 to then choose the suitable cloud services from a single provider to satisfy the Big Data task requirements, and 3) implement multiple providers selection to better satisfy requirements of Big Data workflow composed of multiples tasks. To cope with the multi-criteria selection problem, we extended the Analytic Hierarchy Process (AHP) to better provide more accurate rankings. We develop a set of experimental scenarios to evaluate our 3-phase selection schemes while verifying key properties such as scalability and selection accuracy. We also compared our selection approach to well-known selection schemes in the literature. The obtained results demonstrate that our approach perform very well compared to the other approaches and efficiently select the most suitable cloud services that guarantee Big Data tasks and workflow QoS requirements.

1. Introduction

Cloud Computing is a promising potential venue for processing Big Data tasks as it provides on-demand resources for managing and delivering efficient computation, storage, and cost-effective services. However, managing and handling Big Data implicates many challenges across several levels, among which are the difficulty of handling the dynamicity of the environment resources, the dataflow control throughout the service compositions, and guaranteeing functional and performance quality. Therefore, abundant Cloud Service Providers (CSPs) offering comparable services and functionalities proliferate in the market to meet the growing challenging demands. Subsequently, the selection of the most appropriate cloud provider is recognized to be a challenging task for users. Not only appropriate in terms of functionality provisioned, but also satisfying properties required by the user such as specific levels of quality of service and reputation, especially with the exaggerated cloud providers' marketing claims of guaranteed QoS levels.

Hence, providing an automatic and modest means for selecting a cloud provider which will enable Big Data tasks and guarantee a high level of Quality of Cloud Service (QoCS) is a necessity. Moreover, modeling and evaluation of trust among competing cloud providers enables wider, safer and more efficient use of Cloud Computing.

Therefore, it is necessary to propose a comprehensive, adaptive and dynamic trust model to assess the cloud provider Quality of Service prior to making selection decisions.

*Mohamed Adel Serhani, Email: serhanim@uaeu.ac.ae

A large number of CSPs are available today. Most pf CSPs offers a myriad of services, for instance, Amazon Web Service (AWS) offers 674 varying services which are classified according to locality, Quality Of Service, and cost [1]. Automating the service selection to not only rely of simple criterion such as cost, availability, and processing power, but to consider service quality agreement is crucial. Current CPS selection approaches support straightforward monitoring schemes and do not provide a comprehensive ranking and selection mechanism. For instance, CloudHarmony [2] supports up-to-date benchmark results that do not consider the price while Cloudorado [3] supports price measurement, however neglects other dynamic QoS properties.

Selecting the best CSP reveals twofold objectives, and adds value to both CSPs and Big Data users as well as applications. CSPs provision services that attract clients' interest and support their processing and storage needs. However, users must ensure that services they were offered meet their expectation in terms of quality and price.

Difficulties linked to CSP selection to handle Big Data tasks include for example the following: 1) The limited support for Big Data users in describing their various QoS needs of different Big Data tasks. 2) The difficulty to search in a high dimensional database or repository of CSPs. 3) The challenge to consider the continuous variations in the QoS needs and the Big Data related requirements. And 4) The limited support for mapping Big Data task quality requirements to the underlying cloud services and resources quality characteristics. By doing so, we can guarantee an end-to-end quality support from the top-down Big Data quality consideration to cloud services and resources quality enforcement.

Our main objective in this work is to build a full-fledged approach that supports Big Data value chain with the best cloud services and resources that are trustworthy, automatically scale, and support complex and varying Big Data quality requirements. This is possible with the development of a comprehensive cloud services selection model that fulfills the needs of a Big Data job with the efficient supporting cloud services. Our solution will impose QoS of Big Data processes through dynamic provisioning of cloud services by one or multiple CSPs that will ensure high quality cloud services and fulfill crucial Big Data needs. we propose in this paper a selection approach which includes three phases as follows: our first selection scheme, eliminates CSPs that cannot support the QoS requirements of a Big Data job, which decreases the next selection stage search scope. Consecutively, our second selection stage extends the Analytic Hierarchy Process (AHP) approach to provide selection based on ranking cloud services using various attributes such as Big Data job characteristics, Big Data task profile (BDTP), Quality of Service and considering the continuous changes in cloud services and resources.

The third phase consists of selecting cloud services among different cloud providers, this happens mainly if none of the cloud providers can support the BDTP solely. In addition, if the Big Data job is possibly split into smaller jobs, during the three selection phases, our approach maps the upper quality requirements of the Big Data job to lower level matching quality characteristics of cloud services.

2. Related Work

Cloud service selection attracted the attention of researchers because of its crucial role in satisfying both the users' and providers' objectives having high quality service while optimizing resource allocation and costs. They proposed various approaches to handle and manage the cloud service selection problem. In this section we outline and classify these approaches and emphasize on their strengths and weaknesses.

A broker-based system is described in [4] where the authors proposed a multi-attribute negotiation to select services for the cloud consumer. The quality data is collected during predefined intervals and analyzed to detect any quality degradation, thus allowing the service provider to allocate additional resources if needed to satisfy the SLA requirements. Another broker-based framework was proposed to monitor SLAs of federated clouds [5] with monitored quality attributes measured periodically and checked against defined thresholds. Additionally, in [6], the authors proposed a centralized broker with a single portal for cloud services, CSP, and cloud service users. The authors in [7] proposed a distributed service composition framework for mobile applications. The framework is adaptive, context-aware and considers user's QoS preferences. However, this framework is not suitable for for cloud service selection due to heterogeneity and dynamicity nature of the cloud environments.

The authors in [8] proposed a broker–based cloud service selection framework which uses an ontology for web service semantic descriptions named OWL-S [9]. In this framework, services are ranked based on a defined scoring methodology. First, the services are described using logic-based rules expressing complex constraints to be matched to a group of broker services. Another service selection system was proposed in [10] where the authors proposed a declarative ontology-based recommendation system called '*CloudRecommender'* that maps the user requirements and service configuration. The objective of the system is to automate the service selection process, and a prototype was tested with real-world cloud providers Amazon, Azure, and GoGrid, which demonstrated the feasibility of the system.

In [11], a declarative web service composition system using tools to build state charts, data conversion rules, and provider selection policies was proposed. The system also facilitates translation of specifications to XML files to allow de-centralized service composition using peer-to-peer inter-connected software components. In addition, the authors in [12] proposed a storage service selection system based on an XML schema to describe the capabilities, such as features and performance.

Optimizing the performance is a significant issue in Cloud Computing environments. In other words, better resource consumption and enhanced application performance will be achieved when embracing the appropriate optimization techniques [13]. For example, minimizing the cost or maximizing one or more performance quality attributes. In [14], a formal model was proposed for cloud service selection where the objective is to not only the cost but also the risks (e.g., cost of coordination and cost of maintenance). In this evaluation, the model studies different cost factors, such as coordination, IT service, maintenance, and risk taking. Furthermore, the risks are denoted in terms of integrity, confidentiality, and availability.

The authors in [15] proposed a QoS-aware cloud service selection to provide SaaS developers with the optimized set of

composed services to attend multiple users having different QoS level requirements. They used cost, response time, availability, and throughput as different QoS attributes. The ranking of services is evaluated using integer programming, skyline, and a greedy algorithm providing a near-optimal solution.

Different optimization techniques were adopted for cloud service selection in the literature. One of which were proposed in [16], which used a probabilistic and Bayesian network model. The authors modeled the discovery of cloud service as a directed acyclic graph DAG to represent the various entities in the system. In [18], the authors model cloud service selection as a multi-objective p-median problem according to pre-defined optimization objectives. Their objectives are to optimize the QoS, the number of provisioned services, the service costs, and network transmission costs simultaneously in the given continuous periods. The model also supports the dynamic changing users' requirements over time. Similarly in [17], the authors suggested a service selection model based on combining fuzzy-set multiple attribute decision making and VIKOR. Nevertheless, the discrepancies among user requirements and the providers were not addressed.

The authors in [19] incorporated the IaaS, PaaS, and SaaS service subjective quality attributes based on user preference and applied fuzzy rules based on training samples for evaluation of cloud services quality. A resource management framework is proposed in [20] using a feedback fuzzy logic controller for QoS-based resource management to dynamically adapt to workload needs and abide by SLA constraints. Also, fuzzy logic was adopted in [21] to allow for a qualitative specification of elasticity rules in cloud-based software for autonomic resource provisioning during application execution. A CSP ranking model was proposed in [22] based on user experience, and service quality using an intuitionistic fuzzy group decision making for both quantifiable and non-quantifiable quality attributes to help users select the best CSP conferring to their requirements.

Another cloud service recommendation system was presented in [23] with a selection based on similarity and clustering according to user QoS requirements for SaaS, including cost, response time, availability, and throughput. The users are clustered according to their QoS requirements and are ranked based on multiple aggregation QoS utility functions. Their approach is composed of different phases, starting with clustering the customers and identifying the QoS features, then mapping them onto the QoS space of services, clustering the services, ranking them, and finally finding the solution of service composition using Mixed Integer Programming technology.

Additionally, Multiple Criteria Decision Making (MCDM) models and fuzzy synthetic decision were commonly used in combination for service selection. In [24], fuzzy synthetic decision was applied for selecting cloud providers taking into consideration user requirements. Furthermore, the authors in [25] adopted fuzzy-set theory to evaluate cloud providers trust based on quality attributes related to IaaS. Also in [26], the authors proposed a framework for QoS attributes-based cloud service ranking by applying AHP techniques. A case study was presented to evaluate their framework. Yet, this work was limited to using the measurable QoS attributes of CSMIC rather than including the non-measurable QoS criteria as well [17]. Other works used AHP approach for coud service selection, such as in [1], where the authors adopted MCDM method using AHP to select CPs based on real-time IaaS quality of service. Similarly, The authors in [27] distributed cloud resource management based on SLA and QoS attributes. They adopted AHP to cope with the cloud environment changes during the resource selection process. However, both works exhibit the limitation of only considering the QoS of the cloud services as their selection basis.

Web services frequently undergo dynamic changes in the environment such as overloaded resources. Hence, the authors in [28] proposed a multi-dimensional model, named AgFlow, for component services selection according to QoS requirements of price, availability, reliability, and reputation. The model optimizes the composite service QoS required by the user and revises the execution plan conforming with resource performance dynamic changes. The authors in [29] proposed an SLA renegotiation mechanism to support and maintain QoS requirements in cloud-based systems. They use historical monitoring information including service statuses such as availability, performance, and scalability to predict SLA violations.

Few existing cloud federation projects are based on brokering technologies for multi-cloud composed services. Hence, more research needs to be done towards a standardized methodology for handling interoperability and standard interfaces of interconnected clouds [30]. Trustworthiness evaluation models among different cloud providers were proposed and focus on a fully distributed reputation-based trust framework for federated Cloud Computing entities in cloud federation. In this model, trust values are distributed at each cloud allowing them to make service selection independently [31]. Trust modeling was also tackled in federated and interconnected cloud environments where both consumers and different cloud providers need to trust each other to cooperate [32].

The literature is missing a comprehensive selection model that incorporates all cloud service layers, dimensions, and components in a multi-dimensional model that satisfies service selection for such constrained Big Data applications. Additionally, among the several methods used to determine the user's QoS preference, none exhibits the flexibility to make it responsive to the user's point-of-view as well as comprehends the specific characteristics related to Big Data applications. Accordingly, service selection models are to take into consideration the subsequent requirements: 1) Transparency for stakeholders (such as, customers, CPs, and service brokers), 2) Simple interface that is user friendly, easy to, configure, control and integrate 3) Maintainable and self-adapting to service layers, such as, SaaS, IaaS, and PaaS, and 4) Require low communication overhead by using low number and lightweight messages between stakeholders.

We aim in this work to build a complete, flexible, and QoS driven solution to assess different CSPs' services' capabilities of handling various Big Data tasks. Hence, we develop a three-phase cloud service selection scheme that considers the task complexity and the dynamicity of cloud resource and services. The first step in the selection process consists of apprehending required Big Data quality of service, define and endorse these requirements using the proposed Big Data Task Profile (BDTP). It adopts three selection phases to assess in real-time the CPs QoS and their corresponding services and choose only those that match these requirements.

3. Big Data Task Profile

We explain in this section the main elements of our Big Data specification model as depicted in Figure 1. For every different Big Data task, we model the related profile categories. Additionally, we model a set of attributes and characteristics classifications for

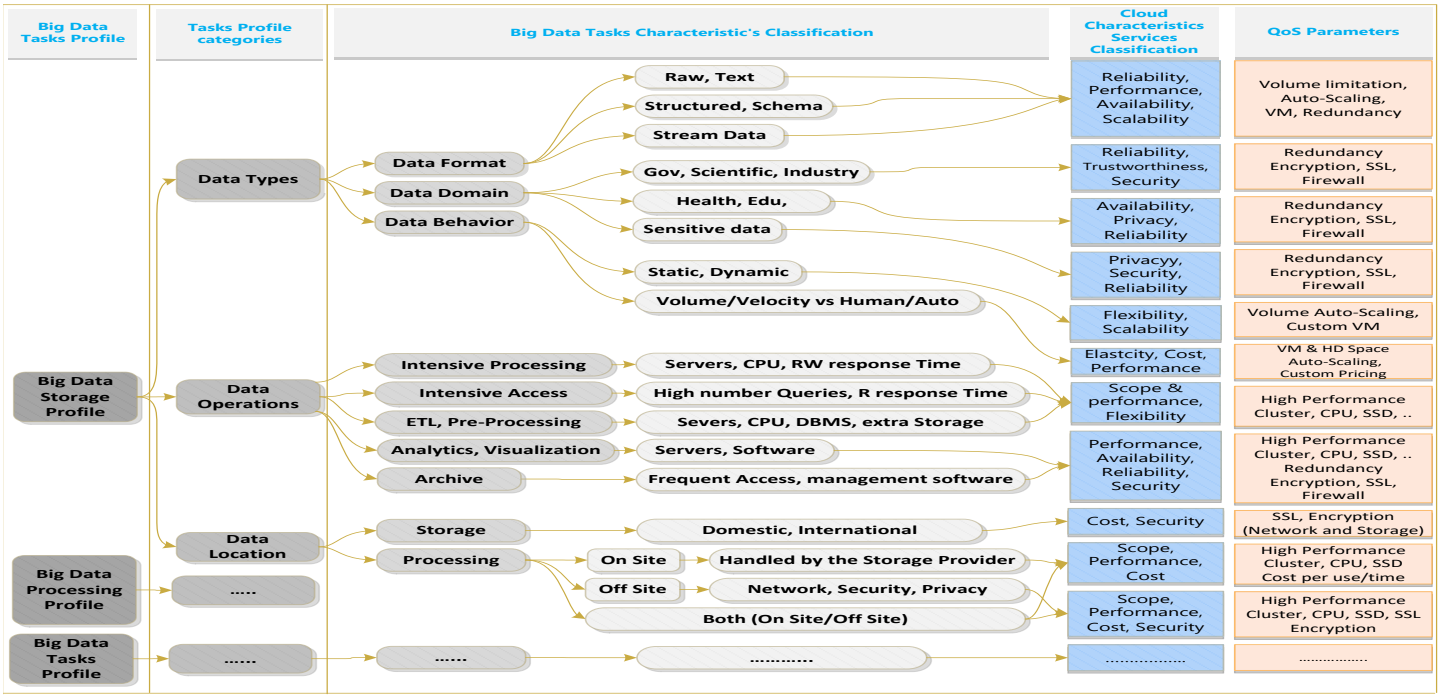


Figure 1. Profile-based Specification Model for BD Tasks

each category. Furthermore, we map the Big Data characteristics to its corresponding cloud attribute and services.

3.1. Big Data Task Profile (BDTP) Specification

The BDTP specifies the main Big Data task requirements that need to be satisfied, and it is modeled as a set of triples: ***R= {DT, DO, DL};*** where , ***DT*** refers to Data Type, and ***DO*** refers to Data Operation, and ***DL*** refers to Data Location. A Big Data request profiled based on BDTP, which defines the requirements and the most appropriate quality specifications that meet a certain Big Data task (such as, Big Data storage). For instance, Storage Profile specifies the following requirements:

Data Types Specifications:

a) Format: structured, unstructured, semi-structured, or stream data.
b) Domain: government, health, smart cities, etc.
c) Behavior: volume static vs dynamic scale, and velocity.

Stored Data Possible Operations:

a) Intensive processing
b) Intensive access
c) Extract Transform Load
d) Analytics and visualization.
e) Archive and backups only

Data Storage Location:

a) Storage Preference
   - Local cloud service provider
   - Geographically disperse site: this involves considering the following properties: network bandwidth, and security of data.

b) Data processing location:
   - On site: security and cost requirements (high or low).
   - Off site: network, security, cost, and servers requirements

Table 1: Sample Profiles for Big Data Tasks

TaBig Data Tasks	Related Cloud Services	Needed Resources	Cloud Services Classification	QoS Parameters
Generation and Collection	PaaS, DaaS	A, C	d, c, e	1, 2
Preprocessing	LaaS	A, B, C, D	a, c, e, f	1, 2, 3, 4
Processing	PaaS, SaaS	A, B, C, D	a, c, e, f	1, 2, 3, 4
Analytics	SaaS	C, D	i, f	1, 3, 4
Visualization	SaaS	C, D	i, f	1, 3, 4
Storage	DaaS	A, C	a, b, c, d, e	1, 2
Transport	LaaS	A	a, b, c, d, e	2
A. Networks B. Servers C. Storage D. Applications E. Infrastructure	1. Storage RW Speed (SSD/HDD) 2. Network Speed Mb/s & Latency 3. CPU Speed, Core, Count 4. RAM size		a. Performance b. Security c. Reliability d. Availability e. Scalability f. Transformation g. Hetriogeneity h. Privacy i. Governance	

Figure 2 illustrates the events issuing succession that deal with a Big Data request. Once a request is received, the best suitable BDTP is selected from the stored profile, in addition, the requirement is normalized to generate a profile *R*. Then the profile is linked with the user's quality of service requirement to produce an updated profile *R'* which will assist in the 3-phase selection. In the first selection stage we generate a list of CSPs ***CPList*** that is used for the second selection phase to generate another list of cloud services *CSList*.

The tipples ***R= {DT, DO, DL}*** represent the BDTP-based user requirements after mapping the appropriate BDTP profile to be

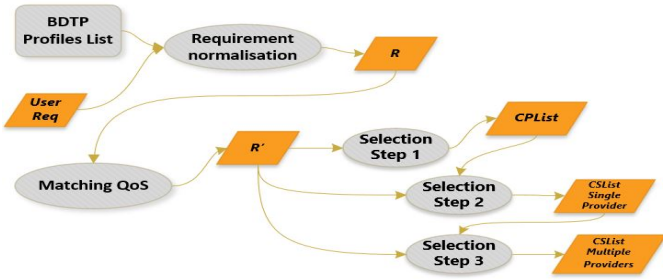


Figure 2. Two-step CSP selection based on BDTP

used. Afterwards, the profiled requirement R are translated into $\boldsymbol{R'= \{QoS_0, QoS_1,\ldots,QoS_i\}}$, for usage in the first selection stage. The later produces a list of CSPs fulfilling $\boldsymbol{R'}$ and noted as: $\boldsymbol{CPList = \{CP_1, CP_2,\ldots,CP_i\}}$. Moreover, the next selection stage based on cloud services is initiated on the same $\boldsymbol{R'}$ to produce a set of cloud services list noted as: $\boldsymbol{CSList = \{CS_1, CS_2,\ldots, CS_i\}}$. The third selection step, retrieves the *CSList* from the second selection phase and look for other cloud services from different providers that can satisfy the request.

3.2. Big Data Workflow Profile (BDTP) Specification

In this section, we describe a simple workflow applied in a case where a patient needs to be continuously monitored to predict epileptic seizures before they actually occur. The monitoring process involves placing multi-channel wireless sensors on the patient's scalp to record EEG signals and continuously stream the sensory data to a smartphone. This process does not restrict the patient's movements. The continuous recorded sensor data, such as 1 GB of data per hour of monitoring is considered a Big Data. However, smartphones lack the capabilities to handle this Big Data, whereas Cloud Computing technologies can efficiently enable acquiring, processing, analyzing, and visualization data generated form monitoring. Figure 3 describes the epilepsy monitoring workflow, where task t_1 is the data acquisition task that is responsible for collecting the EEG data is from the scalp by sensor electrodes then transfers the signals to be preprocessed to computing environment or to temporary storage t_2, which is storing the raw EEG signals. Task t_3 performs data cleansing and filtering processes to eliminate undesirable and noisy signals. Task t_4, is the data analysis task where the EEG data is analyzed to mine meaningful information to provision diagnosis and help decision-making. Finally, t_5 is the task responsible for storing the results.

In this workflow, a task is modeled as a tuple $t\langle tn, in, out\rangle$, where, tn is the task name and in and out are the input and the output data set respectively. Task dependency is modeled in $\boldsymbol{E} = \{(t_i, t_j) | t_i, t_j \in \boldsymbol{T}\}$, where t_j *is* dependent on t_i when t_j is invoked after the t_i is completed. The data flow is modeled by tracking the task input and output states. For each task t_i, we keep information about the data parameters, type and format.

3.3. Matching the BDTP to Cloud Service QoS

As we define $\boldsymbol{R= \{DT, DO, DL\}}$ to be a triple including Data Types, Data Operations and Data Location, we map each request's parameters from high level task specification to a low-level cloud

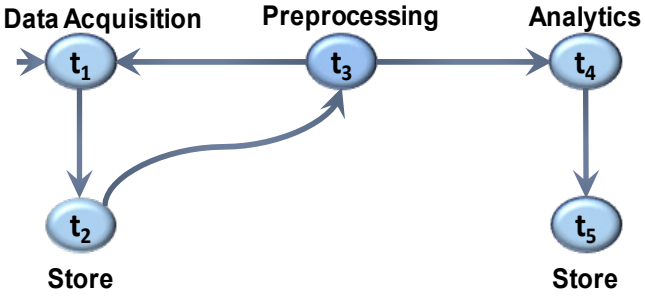


Figure 3. Workflow Example

service's QoS attributes having values and ranges that satisfy each requirement of the BDTP. For each selection phase, the matching process engenders a predefined profile. The QoS Profile is continuously revised to incorporate customer's request needs even after mapping and adjustments of quality attributes. Table 1 illustrates the matching scheme of Big Data tasks to cloud services QoS attributes.

3.4. Web-based Application for Collecting of Big Data Workflow QoS Requirements

In this section, we describe a web-based application we developed for collecting Big Data workflow QoS preferences from the user and generating a quality specification profile, which will be used as basis for task and workflow quality-based trust assessment as shown in Figure 4. This GUI application, collects the quality specification that illustrates the main requirements of a Big Data workflow and its composed tasks. Some of the workflow quality requirements are application domain, data type, operations and location. Furthermore, the application collects the required quality information for every composed task in the workflow, such as quality dimension, quality attributes and the weight values required for the overall trust score calculation. In addition, output data quality is specified for each task along with the weights preferred by the user. Finally, a complete workflow quality profile is generated that enumerates the most suitable requirements and specifications, which fits each Big Data task, such as Big Data preprocessing.

4. Cloud Service Selection Problem Formulation

One of the multi-criteria decision making methods is the Analytic Hierarchy Process (AHP) which is often used for such problems. It adopts a pairwise comparison approach that generates a preferences set mapped to different alternatives [33]. The advantage of AHP methodology is that it allows converting the subjective properties into objective measurements so they can be included in the decision-making, and hence permits the aggregation of numerical measurements and non-numerical evaluation. Additionally, it integrates the user's preference through getting the relative importance of the attributes (criteria) according to the user perception [1]. Accordingly, the quality attributes are represented as a hierarchal relationship, that matches the decision makers form of thinking [34]. Our recommended cloud service selection hierarchy is shown in Figure 5. This hierarchy clearly fits the mapping structure of Big Data to cloud services.

The AHP is intended to pairwise compare all different alternatives which are the quality attributes in our case. Therefore, the more quality attributes are considered, the larger the comparison matrix becomes and the higher number of comparison

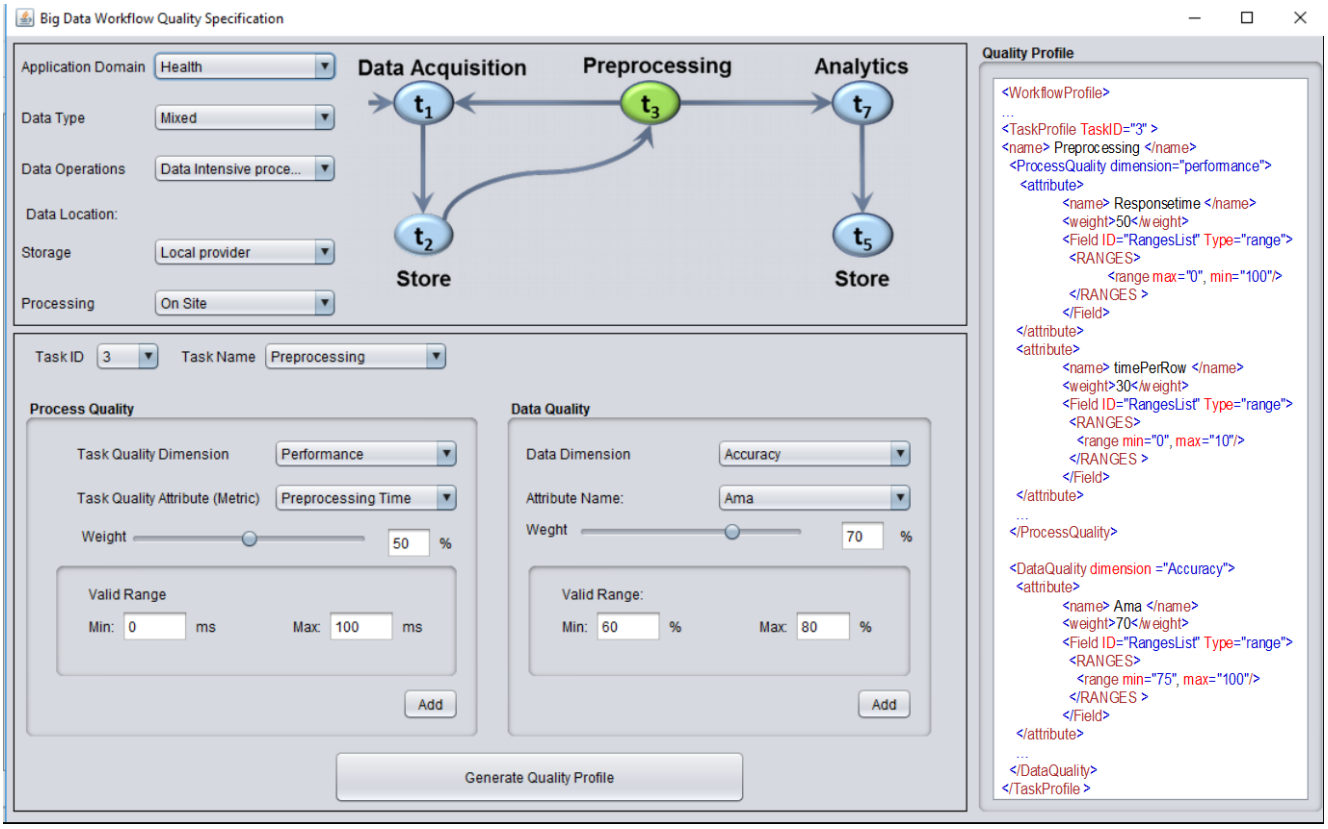


Figure 4. User interface for the collection of Big Data QoCS requirements

will be performed. Hence, we suggest to modify the original AHP approach as in [13].

The idea is to simplify the techniques to avoid the pairwise comparison by normalizing the quality attributes comparison matrix using geometric means which will decrease the required processing to reach a selection decision. Nevertheless, this modification will result in a converged weight matrix as a reason for adopting the geometric mean normalization and hence having a close attribute weight values. Eventually, the attribute priorities will diminish and will not satisfy the objective of this method. To solve this problem, we propose using the simple mean instead of geometric mean for normalization and calculating the attribute weights that matches the user priorities. We followed three steps in our selection approach given as:

4.1. Step1: Hierarchy Model Construction

We adopt the following definitions in our selection model [35]:

Definition 1: *The goal of decision problem which is the main objective and motivation. Here, the goal is the cloud service selection that best matches Big Data task profile conferring to the customer preference.*

Definition 2: *The alternatives which are represented with a set of various options open to the users to be considered in the decision. In our case, they are the group of available cloud services supplied by different cloud providers and matches the recommended BDTP.*

Definition 3: *The criteria of a decision problem. In this case, they are the quality attributes evaluated and upon which the comparison between alternatives is based on to eventually reach a decision. Specifically, they are the QoS attributes provided by the BDTP. The cloud services (alternatives) will be evaluated in comparison to the quality attributes (criteria) for measuring the matching level of the goal of the problem.*

The QoS attributes (criteria) for our decision-making problem are depicted in Figure 1 where they are quantified and qualified using the BDTP by assigning acceptance threshold values or ranges of values [35].

$$\mathrm{CP} = \{cp_i | i = 1,2,3, \dots n\} \tag{1}$$

$$\mathrm{S} = \{s_i | i = 1,2,3, \dots n\} \tag{2}$$

where $\forall\ s_i \in S$ is offered by one $cp_i \in CP$

$$\mathrm{A} = \{a_j | j = 1,2,3, \dots m\} \tag{3}$$

$$\mathrm{P} = \begin{bmatrix} p_{11} & \cdots & p_{n1} \\ \vdots & \ddots & \vdots \\ p_{1m} & \cdots & p_{\mathrm{nm}} \end{bmatrix} \tag{4}$$

where s_1, s_2 … s_n are the existing n alternative cloud services provided to the user. These services may be offered by various providers. a_1, a_2,..., a_m are the QoS attributes (criteria) from the BDTP mapped to the Big Data task required, for example: storage size, processing power, speed, availability, and reliability. p_{ij} is the performance of the ith alternative s with respect to the jth attribute.

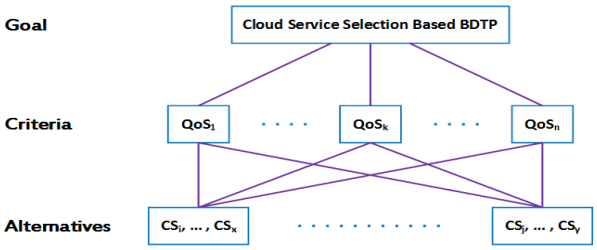


Figure 5. Cloud service selection hierarchy model

4.2. Step2: Attributes Weights and Ranking

AHP scheme consists of mapping each property to a rank or a priority level compared to other criteria applied in different evaluations. Then, an importance level is given by a user for each property opposed to all others [35]. This is performed after building a pairwise comparison matrix using a weighbridge of level of importance. An attribute can be compared to itself and the related importance is set to 1. Therefore, the matrix diagonals are all set to 1 [34]. The importance level is within the range between 1 to 9, where 1 refers to the lowest importance attribute and 9 refers to the most important attribute having the highest value.

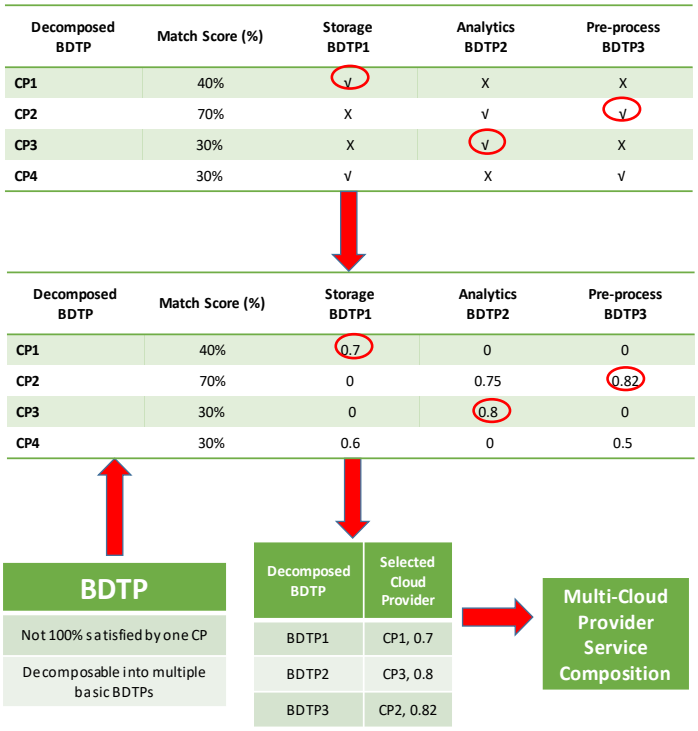
For m attributes, our pairwise comparison of attribute i with attribute j we get a square matrix $A_{M \times M}$ where r_{ij} designates the comparative importance of attribute i with respect to attribute j. This matrix has diagonal values assigned to 1. s.t. $r_{ij} = 1$ when $i = j$. Moreover, it contains reciprocal values across the diagonal, the ratio is inverted s.t. $r_{ji} = 1/r_{ij}$.

$$R = \begin{bmatrix} 1 & r_{12} & \dots r_{1n} \\ 1/r_{12} & 1 & \dots r_{2n} \\ \vdots & \vdots & 1 \end{bmatrix} \quad (5)$$

Then, we define a normalized weight w_i for each attribute based on the geometric mean of the ith row. We choose the geometric mean methodology as an extended version of AHP for its simplicity, easiness of calculating the maximum Eigen value, and for decreasing the inconsistencies of judgment using $GM_i = \left(\prod_{j=1}^{m} r_{ij}\right)^{1/m}$ [34]. After that, the geometric means are normalized for all rows in the matrix using $w_i = GM_i / \sum_{i=1}^{m} GM_i$. Nevertheless, we get equal weights which disallow differentiation between attributes importance. Thus, we suggest to apply the normalized mean values for each row as follows:

$$M_i = \sum_{j=1}^{m} r_{ij} \quad (6)$$

$$w_i = M_i / \sum_{i=1}^{m} M_i \quad (7)$$



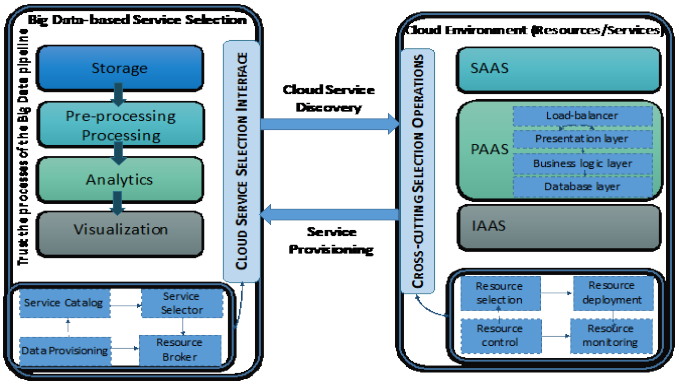


Figure 6. Big Data workflow quality enforcement based on cloud service selection

Table 2: Multi-provider cloud service selection m-pcss

4.1. Step 3: Calculate the Ranking Score of All Alternatives

To generate the rating scores for each cloud service (alternative), we use Simple Additive Weighting method by multiplying weights obtained from eq. 7 w_j of each attribute j with its corresponding performance value in Matrix P from eq. 4. Then summing all resulted values as in:

$$Score_i = \sum_{j=1}^{m} w_j \times (m_{ij})_{normal} \quad (8)$$

Where $(m_{ij})_{normal}$ is the normalized value of m_{ij} and $Score_i$ is the overall rating score of the alternative cloud service S_i. Finally, we select the cloud service (alternative) that has the highest score value:

$$S_{bestscore} = \max_{1 \le i \le n} Score_i \quad (9)$$

5. Model for Cloud Service Selection

We here describe our cloud service selection model to fulfill the quality of Big Data workflow over federated clouds. Figure 6 overviews how various Big Data processes, including storage, processing, and analytics can be provisioned with the cloud services and resources efficiently and with high quality. It details the main components involved in cloud service discovery and provisioning for Big Data value chain. Such components used for selection include service catalog, service broker, and service selector. However, components involved in cloud service provisioning in response to cloud service selection requests include resource selection, deployment, control, and monitoring.

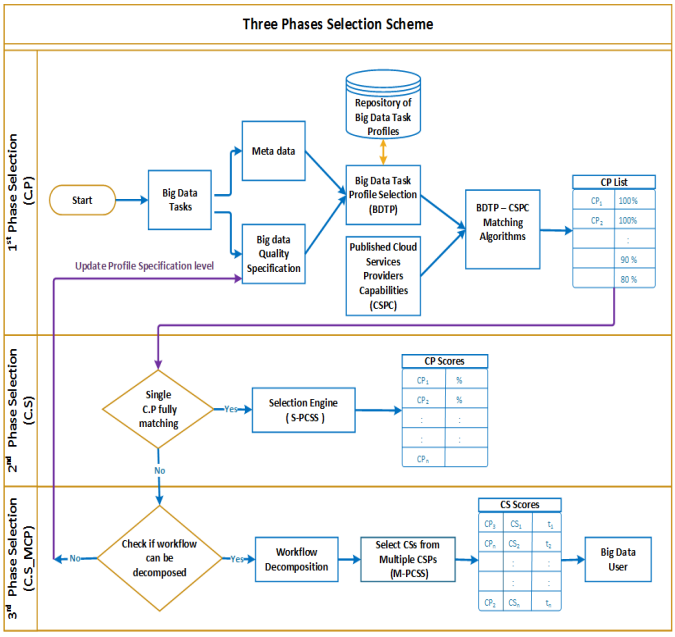


Figure 7. Three phases cloud service selection model

5.1. Cloud Service Selection

As soon as a service request is issued to support Big Data processing and storage while guaranteeing certain QoS, cloud resources are reserved to deploy and process Big Data workflow over the cloud infrastructure. Then, the workflow execution is monitored to detect if any performance degradation occurred and respond with the appropriate adaptation actions to maintain high quality service provisioning.

Figure 7, describes the selection scheme which is implemented in three phases: the first phase involves choosing the most suitable CSPs that conform to the Big Data workflow requirements, however the second phase involves choosing among CSPs the services that fulfill the Big Data Task profile (BDTP). The third phase selection consists of conducting further selection strategy to choose services from different CSPs that satisfy different tasks of a single workflow and maximize the overall quality of the workflow. In the following, we describe in detail each of the three selection phases:

CSP selection phase: Big Data workflows described as an aggregation of tasks present a set of quality requirements, such as, trust, in addition, to extra information known as metadata, such as, type of data, and its characteristics. The Big Data task profile selection component takes as input the metadata and the Big Data quality specification to find and retrieve the closest suitable profile from the Big Data profile repository that responds to the task(s) quality requirements. Both selected profile and published cloud provider's competencies are used to trigger the execution of the CP-Profile matching algorithm which matches the BDTP profile to the CSP published competencies. A list containing scored CSPs is generated by this algorithm. A score granted to each provider refers the ratio of which the CSP is capable to accomplish the Big Data task(s) given the set of quality requirements.

CS selection phase with single provider: the second selection phase is initiated to choose the corresponding cloud services from the list of phase 1 selected CSPs according to two stages:

Stage 1: A single provider cloud service selection algorithm (S_PCSS) is performed if a specific cloud provider completely matches the QoS of the Big Data task. The output of this algorithm is a list of CSPs with their measured scores. Here, we provide an extension of the AHP Method to use a simple mean instead of geometric mean to measure the attribute weight. This leads to variation in the generated weight values for each attribute that matches the pairwise importance levels given by the user.

Stage 2: A process of decomposing Big Data workflow into tasks is triggered if no single CSP is able to fulfil the QoS of the BDTP. Tasks of the workflow should be independent and can be processed impartially. If a workflow cannot be decomposed into undependably executable tasks, a loopback to previous phase will allow reviewing the profile specification to meet the selection measures.

CS selection phase with multiple providers: the third selection phase. Once a workflow can be decomposed into a set of tasks, the multi-provider cloud service selection algorithm is implemented to cope with multiple service selection from various cloud providers to maintain the quality of aggregated workflow tasks. Table II depicts an example of BDTP decomposition into three independent profiles for storage, pre-processing, and analytics. A score is calculated for each CSP with regards to each profile and cloud providers that have the highest score are selected to handle each profile independently.

5.2. Selection Algorithms

According to the scheme described in Figure 7, we have developed three consecutive algorithms to support the three phases selection as follows:

The BDTP-CSPC algorithm: maps the BDTP with each CSP Capabilities (CSPC), for example, availability and cost. The selection is performed according to the providers' capabilities satisfaction without considering customer favoured priorities. Figure 8 describes the algorithm which requires the list of CSPs, the list of required quality attributes (profile) and the list of published quality attributes for each cloud provider. Then performs one-to-one matching of each pair of attributes (profile-published) and outputs a list of scored CSPs which completely match the BDTP. Each CSP is linked to a set of provided quality characteristics. The algorithm performs an evaluation of each CSP matching score based on the percentage of fulfilled quality attributes required by the BDTP. The BDTP-CSPC matching

Algorithm 1 BDTP-CSPC matching algorithm

Input: $CPList$ //List of CPs,
$profileAttrList$ //List of required quality attributes (profile),
$attrListPerCP$ //List of published quality attributes of each cloud provider

Output: $cpScoresList$ //Matching Score for each CP

```
1: procedure CPMatchScore(CPList, attrListPerCP, profileAttrList)
2:     for all cp ∈ CPList do
3:         for all attr ∈ profileAttrList do
4:             cpAttrValue ← find(attr, attrListPerCP[cp])
5:             if (matchRequirement (cpAttrValue, attr))
6:                 count ← count + 1
7:             endif
8:             score[cpScoresList] ← count/size(profileAttrList)
9:         end for
10:    end for
11:    return cpScoresList[cp]
12: end procedure=0
```

Figure 8. BDTP-CSPC matching algorithm

Algorithm 2 Single Provider Cloud Service Selection (S_PCSS) algorithm

Input: *serviceList*, //List of CSs
profileAttrList, //List of required quality attributes (BDTP)
srvcAttrValue, //List of published quality attributes of each CS

Output: *srvcScoresList* //Matching Score for each CS

```
procedure EvaluateServiceAHP_S_CP(serviceList, profileAttrList, srvcAttrValue)
    R ← buildComparativeImportanceMatrix(BDTP, profileAttrList)
    M ← R.size()
    sumMN ← 0
    for s_i ← 1, M do // each service (row)
        MN_i ← 1
        for attr_j ← 1, M do // each attribute (col)
            MN[s_i] ← MN[s_i] + R[s_i][attr_j]
        end for
        sumMN = sumMN + MN[s_i]
    end for
    for i ← 1, M do // each service (row)
        weights[i] = MN[i]/sumMN // weight for each quality attribute
    end for
    for i ← 1, serviceList.size() do
        score_i ← 0
        for j ← 1, srvcAttrValue.size() do
            score_i ← score_i + srvcAttrValue[i][j] * weights[j]
        end for
        srvcScoresList[i] ← score_i
    end for
end procedure=0
```

Figure 9. Cloud Service Slection Algorithm - Single Provider

algorithm is scalable with the proliferation of cloud providers and quality attributes considered.

The S_PCSS algorithm: handles the second stage selection mechanism that considers thorough information about the attributes described in the BDTP to provide ranking values of the cloud services offered by the selected CSPs by the BDTP-CSPC algorithm. We adopted AHP and MADM to implement our selection strategy of cloud services. Figure 9 explains the single selection algorithm that uses a list of cloud services, the list of required quality attributes (BDTP), and the list of published quality attributes for each cloud service. Then, it generates a comparative matrix identifying the priority level of each published quality attribute in comparison to other quality attributes existing in the BDTP. Afterwards, this matrix is used to calculate and return a list of ranked cloud services with the highest scores and satisfy the Big Data task profile.

Algorithm 3 Multi-Provider Cloud Service Selection (M_PCSS) algorithm

Input: *cpList*, //List of CPs
serviceList, //List of CSs
profileAttrList, //List of required quality attributes (BDTP)
srvcAttrValue, //List of published quality attributes of each CS
cpScoresList, //List of CPs scores

Output: *srvcScoresList* //Matching Score for each CS

```
procedure EvaluateServiceAHP_M_CP(cpList, serviceList, profileAttrList, srvcAttrValue, cpScoresList)
    for cp_i ← 1, N do // each cloud service provider
        cpServiceScore[cp_i] ← EvaluateServiceAHP_S_CP(serviceList, profileAttrList, srvcAttrValue[cp_i]))
        for attr_j ← 1, M do
            cpServiceScore[cp_i] ← cpServiceScore[cp_i] × cpScoresList[cp_i]
        end for
    end for
    for attr_j ← 1, M do // each attribute (col)
        maxAttr ← 1
        for cp_i ← 2, N do // each cp (row)
            if (cpServiceScore[cp_i][attr_j] ≥ cpServiceScore[maxAttr][attr_j])
                maxAttr ← cp_i
        end for
        srvcScoresList[attr_j] ← tuple(maxAttr, cpServiceScore[maxAttr][attr_j
    end for
    return srvcScoresList
end procedure=0
```

Figure 10. Cloud Service Slection Algorithm - Multi- Provider

The M_PCSS algorithm: this algorithm handles the third stage selection where none of the CSPs fully supporting the Big Data workflow. In this situation, the workflow is decomposed into single independent tasks which will be processed by different cloud providers. Figure 10 describes the M_PCSS algorithm, the later takes as input the list of cloud providers, their offered cloud services and their calculated scores as well as the list of required quality attributes (BDTP), and the list of published quality attributes for each cloud service. It first applies the ***S_PCSS*** algorithm to receive the cloud service scores within each cloud provider. Then it finds the best matching services having the highest score among all cloud providers. Additionally, this algorithm favors the cloud provider that provides more services to minimize the communication and cost overhead due to data transfer and processing distribution. This is achieved by multiplying the cloud provider score to the service score to reach a final cloud service score.

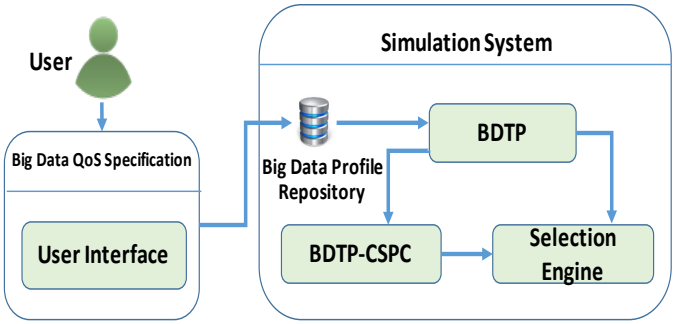


Figure 11. Simulator Components

6. Evaluation of Cloud Service Selection

This section details the experiments we conducted to assess the three-phase selection approach using various experimental scenarios.

6.1. Environment Setting

The setting and the simulation parameters we have used to conduct the experiments are described hereafter:

Setting and simulation parameters
Desktop: CPU Intel Core TM i7-3770K @ 3.40 GHz and Turbo Boost, DDR3 RAM 32GB, HD 1TB, and OS 64-bit.
Number of CSPs: 1 - 100.
Number of services provided by each CSP: 1 - 100.
QoS attributes: data size, distance, cost, response time, availability, and scalability.

6.2. Simulator

Figure 11 depicts the main modules of the JAVA simulator we have developed to implement the selection algorithms we have developed to support the three selection phases of cloud service providers and their related cloud services based on the BDTP and

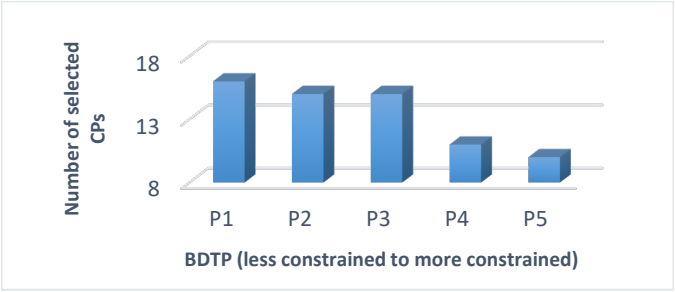


Figure 12. CSP profile-based matching with different constrains levels

the AHP method. The simulator comprises five main components as follows:

BDTP component: this module classifies the Big Data task requests into three categories: data type, data operation and data location. It also sets the acceptance level (minimum, maximum, threshold), for each quality property and eventually normalizes the performance scores.

BDTP-CSPC component: integrates the full implementation of BDTP-CSPC selection algorithm we described above. This module measures a score for each cloud provider that matches the BDTP. CSPs scoring 100% are nominated to the second phase selection Engine.

Selection Engine: integrates the implementation of the S_PCSS algorithm. The later uses the BDTP and the selected CSPs nominated in the first phase, then implements AHP to rank and retrieve the set of cloud services from the list of CSPs that fulfil Big Data task. Moreover, the selection engine implements the M_PCSS selection algorithm to incorporate the implementation of selecting cloud services from different CSPs while calculating cloud services scores for each cloud provider. Afterwards, it selects the best matching cloud service with the highest score among all cloud providers.

Big Data QoS specification: it supports and guides users through an interface to specify the Big Data task quality attributes as depicted in Figure 4 above.

Big Data profile repository: serves as repository of Big Data task profiles. It is accessed to retrieve the appropriate profile when a Big Data task request is issued and a selection of suitable CSP and services need to take place to respond to the initiated request.

In addition, to the above implemented entities, the simulator generates multiple CSPs offering multiple cloud services having various QoS attributes performance levels to produce a CSP list that serves the selection algorithms. Other implemented modules include, communication interfaces, scoring schemes implementation, invocation interfaces, and storage management interfaces.

6.3. Experimental Scenarios

In this sub-section, we detail the various scenarios we have chosen to assess our 3-phase selection model and the related implemented algorithms. Scenarios were selected to validate three main properties: CSP selection accuracy, model scalability, and communication overhead.

In the following, we explain the developed scenarios to help evaluating our 3-Phase selection model.

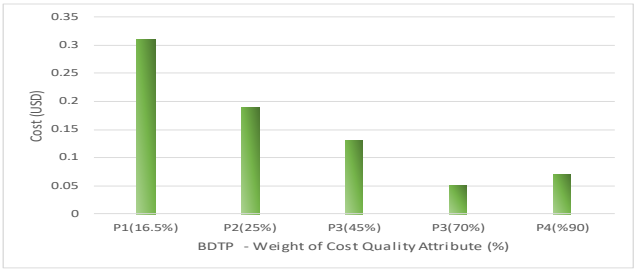


Figure 13. CSP selected with different levels of BDTP strictness (Cost)



Figure 14. CSP selected with different levels of BDTP strictness (Response Time)

Scenario 1: evaluates the accuracy of the the first phase selection in terms of retrieving different Big Data task profiles while fixing the number of cloud providers to 20 CSPs. Figure 12 demonstrates that the less the number of selected CSPs the more the BDTP becomes constrained (e.g. includes extensive quality constraint to consider and evaluate).

Scenario 2: evaluates the accuracy of the the second phase selection based AHP while varying profiles and fixing the number of cloud providers. This will also retroactively validate the first selection results. Figure 13, demonstrates that the more constrained the BDTP is, which will add more weight on the cost quality attribute, the more the recommended CS provides a better cost. In the same manner, Figure 14, stresses the same results but

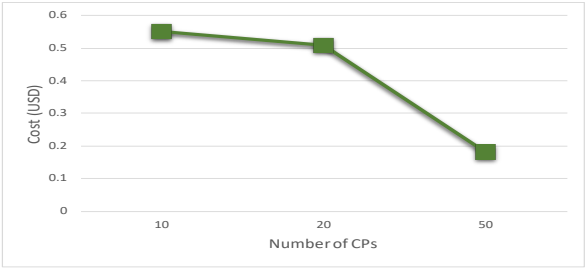


Figure 15. Cost of selected CSP with an increasing number of available CSPs

now with the response time quality attribute.

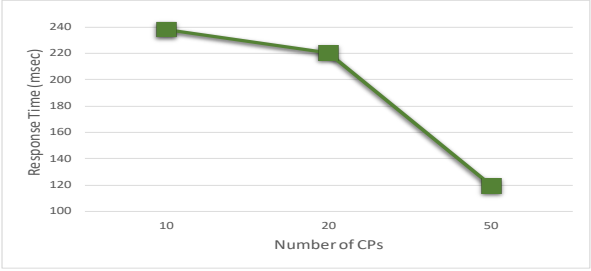


Figure 16. Response Time of selected CSP with an increasing number of available CSPs

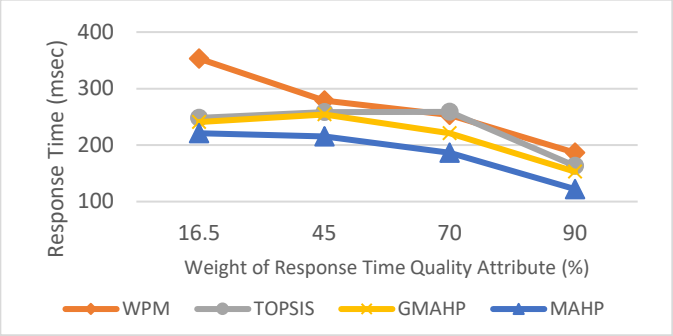


Figure 17. MAHP model behavior with different QoS attribute weight values benchmarked with other models.

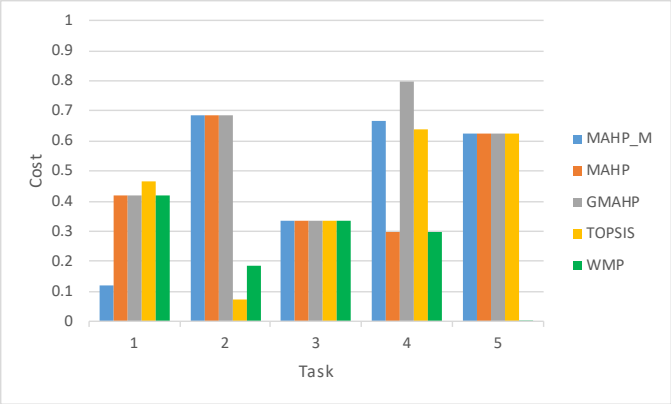


Figure 18. Cost of selected task per algorithm.

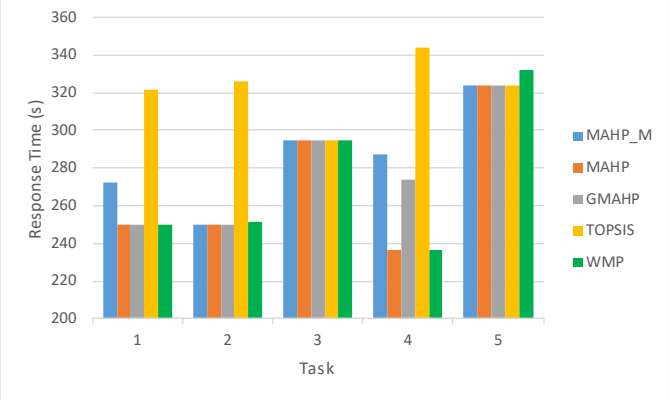


Figure 19. Response time of selected task per algorithm.

Scenario 3: evaluates the scalability of the selection model while increasing the number of CSPs and measuring the QoS cost and response time of the selected CSs after executing the 3-Phase selections respectively. Figure 15 and Figure 16, demonstrate that our 3-phase selection scheme scales perfectly as elucidated through a decrease in the cost and the response time respectively as the number of cloud providers increase. This is because more options are available to select among them which leads to better QoS fulfilment.

Scenario 4: we compare our 3-phase selection scheme with other MADM selection schemes. Our model used the *simplified* AHP using mean values of pairwise comparison matrix (MAHP). However, the other models used are Weighted Product Method (WPM), Technique for Order Preference by Similarity to Ideal

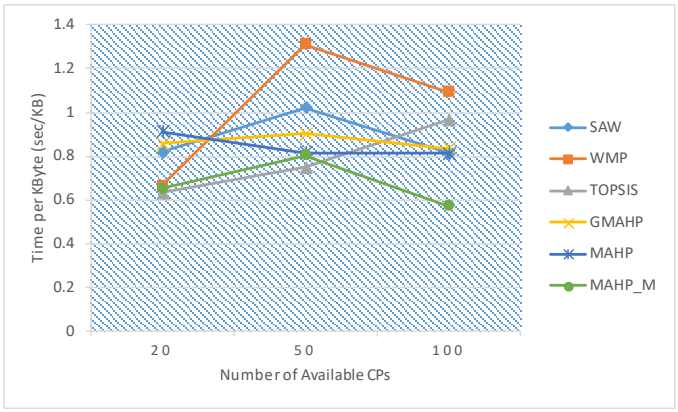


Figure 20. Average response time of selected services using (MAHP) and (MAHP_M) benchmarked with other models.

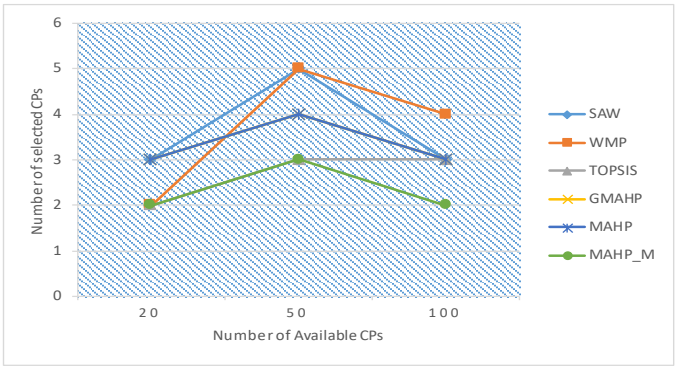


Figure 21. Number of differnet CSPs using (MAHP) and (MAHP_M) benchmarked with other models .

Solution (TOPSIS), and the modified Geometric Means AHP (GMAHP) to make cloud service selection. Comparison was conducted based on the response time quality attribute. Figure 17 demonstrates that MAHP gives better results compared to all other models, it provisions lower response times for all levels of selected quality attribute weights.

Scenario 5: we compare our 3-phase selection algorithm to other MADM selection methods by showing the cost and response time for each task composed in the workflow. As depicted in Figure 18 and Figure 19, the (MAHP) provisions lower task cost and response time respectively, and gives similar results as (GMAHP) and (TOPSIS). However, our *modified* AHP (MAHP_M) method provisions higher cost and response time per task than the (MAHP) since it gives higher preferences to selection of services from an existing cloud provider to minimize the communication and data transfer overhead.

Scenario 6: we compare the average response time of the workflow composed services using different selection schemes. The measured response time includes the communication and transfer overhead due to using services from different CSPs. As depicted in Figure 20, our (MAHP_M) method has the lowest response time because it minimizes the number of CSP among which services are selected, hence, minimizes the time wasted in communication overhead. In addition, Figure 21, shows the number of different CSPs providing the selected services and it shows that our (MAHP_M) method has the lowest number of CSPs and hence has the lowest overhead.

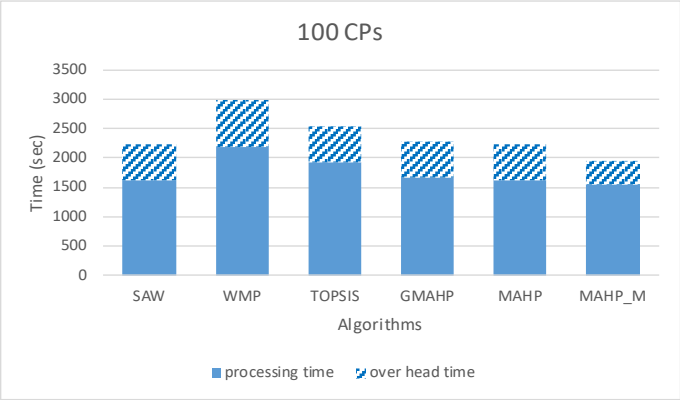


Figure 22. Workflow overhead and processing time of the MAHP and MAHP_M models benchmarked with other models.

Scenario 7: we compare the communication and data transfer overhead due to using different cloud providers. In this scenario, we used 100 CSPs and measured the total workflow execution time and the overhead time when using different selection methods. As shown in Figure 22, our (MAHP_M) method has the least overhead and accordingly total time amongst the rest of the methods. This is because our (MAHP_M) favors services that belong to already selected CSPs to minimize the overhead.

7. Conclusion

Big Data has emerged as a new paradigm for handling gigantic data and get valuable insights out of it. The special characteristics of Big Data reveals new requirements in terms of guaranteeing high performance and high quality of various Big Data processes (e.g. processing, storage, and analytics). The cloud infrastructure and resources are considered a perfect source of resources and services to support Big Data specific quality requirements. Selecting among myriad of cloud service providers the appropriate services and resources that meet these requirements is challenging given the diversity and the complexity of Big Data workflows.

In this paper, we proposed an efficient federated cloud service selection to support workflow Big Data requirements. It is a 3-phase selection scheme which is implemented through three phases. In the first selection phase, it captured the Big Data QoS requirements through the BDTP. However, in the second selection phase, a scored list of cloud services that satisfies the BDTP is generated. Finally, the third selection phase goes further and scored cloud services from different CSPs to better match the workflow quality requirements.

The main contributions of our selection scheme is the integration of a BDTP that ensures the QoS of Big Data tasks and is considered as a reference model for the three successive selection phases. In addition, revising the profile is advisable to have an efficient selection decision. We proposed a further contribution by extending the AHP method by adopting the mean values of pairwise comparison matrix alternative than using the geometric mean. The later shown weakness in producing a weight matrix with equal values of weights for all attributes. The last contribution is supporting workflow key requirements through the selection of multiple cloud services form multiple CSPs which maximized the Big Data complex workflow requirement fulfilment.

We conducted extensive experimentation to evaluate different properties of our 3-phase selection scheme. The results we have obtained proved that our selection model: integrated well the BDTP and guaranteed Big Data QoS requirements, scaled with the growing number of CSPs, performed better than the other MADM schemes such as TOPSIS, WPM, and the SAW, and enforced QoS requirement of Big Data workflows through varying cloud services from multiple CSPs.

For future work, we plan to have an extension for our selection scheme with more scenarios and complex Big Data workflows where other properties such as data security and privacy can also be considered. Furthermore, we are considering to assess our selection scheme against various selection techniques where we use an existing cloud environment.

References

[1] M. Zhang, R. Ranjan, M. Menzel, S. Nepal, P. Strazdins and L. Wang, "A Cloud Infrastructure Service Recommendation System for Optimizing Real-time QoS provisioning Constraints," IEEE Systems Journal, pp. 1-11, 2015. arXiv preprint arXiv:1504.01828

[2] "Cloudharmony," [Online]. Available: http://cloudharmony.com/. [Accessed 5 2017].

[3] "Cloud Computing Comparison Engine," [Online]. Available: http://www.cloudorado.com/. [Accessed 5 2017].

[4] E. Badidi, "A Cloud Service Broker for SLA-based SaaS provisioning," in International Conference on Information Society (i-Society), IEEE, 2013. https://doi.org/10.5121/ijccsa.2016.6201

[5] A. Al Falasi, M. A. Serhani and R. Dssouli, "A Model for Multi-levels SLA Monitoring in Federated," in IEEE 10th International Conference on Autonomic and Trusted Computing (UIC/ATC) Ubiquitous Intelligence and Computing, 2013. https://doi.org/10.1109/uic-atc.2013.14

[6] J. Lee, J. Kim, D.-J. Kang, N. Kim and S. Jung, "Cloud service broker portal: Main entry point for multi-cloud service providers and consumers," in IEEE 16th International Conference on Advanced Communication Technology, 2014. https://doi.org/10.1109/icact.2014.6779131

[7] Q. Z. Sheng, B. Benatallah, Z. Maamar, M. Dumas and A. H. Ngu, "Enabling personalized composition and adaptive provisioning of web services," in International Conference on Advanced Information Systems Engineering, 2004. https://doi.org/10.1007/978-3-540-25975-6_24

[8] L. D. Ngan and R. Kanagasabai, "Owl-s based semantic cloud service broker," in IEEE 19th International Conference on Web Services (ICWS), IEEE, 2012. https://doi.org/10.1109/icws.2012.103

[9] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider and U. Sattler, "OWL 2: The next step for OWL," Web Semantics: Science, Services and Agents on the World Wide We, vol. 6, no. 4, pp. 309-322, 2008. https://doi.org/10.2139/ssrn.3199412

[10] M. Zhang, R. Ranjan, S. Nepal, M. Menzel and A. Haller, "A declarative recommender system for cloud infrastructure services selection," in International Conference on Grid Economics and Business Models, Springer, 2012. https://doi.org/10.1007/978-3-642-35194-5_8

[11] B. Benatallah, M. Dumas, Q. Z. Sheng and A. H. Ngu, "Declarative composition and peer-to-peer provisioning of dynamic web services," in 18th International Conference on Data Engineering. Proceedings., 2002. https://doi.org/10.1109/icde.2002.994738

[12] A. Ruiz-Alvarez and M. Humphrey, "An automated approach to cloud storage service selection," in the 2nd international workshop on Scientific cloud computing, ACM, 2011. https://doi.org/10.1145/1996109.1996117

[13] F. Pop and V. Cristea, "The Art of Scheduling for Big Data Science," in Big Data: Algorithms, Analytics, and Applications, Chapman and Hall, 2015, p. 105–120.

[14] B. Martens and F. Teuteberg, "Decision-making in cloud computing environments: A cost and risk based approach," Information Systems Frontiers, vol. 14, no. 4, pp. 871-893, 2012. https://doi.org/10.1007/s10796-011-9317-x

[15] Q. He, J. Han, Y. Yang, J. Grundy and H. Jin, "QoS-driven service selection for multi-tenant SaaS," in IEEE 5th international conference on Cloud Computing (Cloud), 2012. https://doi.org/10.1109/cloud.2012.125

[16] A. O. Akinwunmi, E. A. Olajubu and G. A. Aderounmu, "A trustworhty model for reliable cloud service discovery," International Journal of Computer Applications, vol. 87, no. 16, 2014. https://doi.org/10.5120/15293-3962

[17] H. M. Alabool and A. K. Mahmood, "Trust -Based Service Selection in Public Cloud Computing Using Fuzzy Modified VIKOR Method," Australian Journal of Basic and Applied Sciences, vol. 7, no. 9, pp. 211-220, 2013. ISSN 1991-8178

[18] H. K. Chen, C. Y. Lin and J. H. Chen, "A multi-objective evolutionary approach for cloud service provider selection problems with dynamic demands," in European Conference on the Applications of Evolutionary Computation, Springer. https://doi.org/10.1007/978-3-662-45523-4_68

[19] S. Telenyk, P. Bidyuk, E. Zharikov and M. Yasochka, "Assessment of cloud service provider quality metrics," in International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), IEEE, 2017. https://doi.org/10.1109/ukrmico.2017.8095422

[20] C. Anglano, M. Canonico and M. Guazzone, "FC2Q: exploiting fuzzy control in server consolidation for cloud applications with SLA constraints," Concurrency and Computation: Practice and Experience, vol. 27, no. 17, pp. 4491-4514, 2015. https://doi.org/10.1002/cpe.3410

[21] P. Jamshidi, A. Ahmad and C. Pahl, "Autonomic resource provisioning for cloud-based software," in In Proceedings of the 9th International Symposium on Software Engineering for Adaptive and Self-Managing Systems, ACM, 2014. https://doi.org/10.1145/2593929.2593940

[22] S. S. Wagle, M. Guzek and P. Bouvry, "Cloud service providers ranking based on service delivery and consumer experience," in IEEE 4th International Conference in Cloud Networking (CloudNet), IEEE, 2015. https://doi.org/10.1109/cloudnet.2015.7335308

[23] Y. Wang, Q. He and Y. Yang, "Qos-aware service recommendation for multi-tenant saas on the cloud," in EEE International Conference on Services Computing (SCC), 2015. https://doi.org/10.1109/scc.2015.33

[24] S. Wang, Z. LiuQibo , S. Zou and F. Yang, "Towards an accurate evaluation of quality of cloud service in service-oriented cloud computing," Journal of Intelligent Manufacturing, vol. 25, no. 2, p. 283–291, 2014. https://doi.org/10.1007/s10845-012-0661-6

[25] M. Alhamad, T. Dillon and E. Chang, "A trust-evaluation metric for cloud applications," International Journal of Machine Learning and Computing, vol. 1, no. 4, pp. 416-421, 2011. https://doi.org/10.7763/ijmlc.2011.v1.62

[26] S. K. Garg, S. Versteeg and R. Buyya, "A framework for ranking of cloud computing services," Future Generation Computer Systems, vol. 29, no. 4, pp. 1012-1023, 2013. https://doi.org/10.1016/j.future.2012.06.006

[27] S. Khaddaj, "Cloud computing: service provisioning and user requirements," in 11th IEEE International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), 2012. https://doi.org/10.1109/dcabes.2012.76

[28] L. Zeng, B. Benatallah, A. H. Ngu, M. Dumas, J. Kalagnanam and H. Chang, "Qos-aware middleware for web services composition," IEEE Transactions on software engineering, vol. 30, no. 5, pp. 311-327, 2004. https://doi.org/10.1109/tse.2004.11

[29] A. F. M. Hani, I. V. Paputungan and M. F. Hassan, "Renegotiation in service level agreement management for a cloud-based system," ACM Computing Surveys (CSUR), vol. 47, no. 3, p. 51, 2015. https://doi.org/10.1007/978-3-642-40861-8_9

[30] A. N. Toosi, R. N. Calheiros and R. Buyya, "Interconnected cloud computing environments: Challenges, taxonomy, and survey," ACM Computing Surveys (CSUR), vol. 47, no. 1, p. 7, 2014. https://doi.org/10.1145/2593512

[31] J. Abawajy, "Determining Service Trustworthiness in Intercloud Computing Environments," in the 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN'09), 2009. https://doi.org/10.1109/i-span.2009.155

[32] D. Bernstein and D. Vij, "Intercloud security considerations," in IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom), 2010. https://doi.org/10.1109/cloudcom.2010.82

[33] M. Whaiduzzaman, A. Gani, N. Badrul Anuar, M. Nazmul Haque and I. Tanzeena Haque, "Cloud service selection using multicriteria decision analysis," The Scientific World Journal, 2014. https://doi.org/10.1155/2014/459375

[34] R. V. Rao, "Improved multiple attribute decision making methods," Decision Making in Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods, pp. 7-39, 2013. https://doi.org/10.1007/978-1-4471-4375-8_2

[35] S. Klutho, "Mathematical Decision Making: An Overview of the Analytic Hierarchy Process," 2013. [Online]. Available: www. whitman. edu/mathematics/SeniorProjectArchive/2013/Klutho. pdf. [Accessed 25 5 2017].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 236-240 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A novel mixed-mode universal biquad employing plus current output DVCCs

Takao Tsukutani*,[1], Yasutomo Kinugasa[2], Noboru Yabuki[3]

[1] *Professor Emeritus, National Institute of Technology, Matsue College, 690-8518, Japan*

[2] *National Institute of Technology, Matsue College, Department of Electrical and Computer Science, 690-8518, Japan*

[3] *National Institute of Technology, Tsuyama College, Department of Integrated Science and Technology, 708-8509, Japan*

ARTICLE INFO

Article history:
Received: 02 June, 2018
Accepted: 28 July, 2018
Online: 05 August, 2018

Keywords:
Analog signal processing
Mixed-mode biquad circuit
Plus current output DVCCs

ABSTRACT

This paper refers to a novel mixed-mode universal biquad employing plus current output differential voltage current conveyors with grounded passive components. The circuit performs mixed-mode operation with the selection of the input and output terminals, and it enables low-pass, band-pass, high-pass, band-stop and all-pass responses choosing suitable input signals. The circuit needs no component matching constraints for obtaining five basic circuit responses above, and can adjust orthogonally the characteristic parameters by the circuit component. Additionally, the circuit has extremely low sensitivities with respect to the circuit components. A circuit design is performed with PSPICE simulation in order to verify the workability of the circuit.

1. Introduction

Active circuit with high performances (i.e. high frequency operation, low power dissipation, wide dynamic range, etc.) is receiving significant attention. Numerous circuit designs using some active devices such as operational trans-conductance amplifiers (OTAs), second generation current conveyors (CCIIs) and differential voltage current conveyors (DVCCs), etc. have been reported in the literature earlier [1-12].

A DVCC is a very useful active device, and DVCC-based circuit is adaptable to wideband operation. A plus current output DVCC (PO-DVCC) is composed of simpler circuit configuration than a minus current output one. Hence it has low power performance compared to the minus current output DVCC. Several voltage-mode (VM) and current-mode (CM) circuits have been synthesized by employing the DVCCs previously [8-11].

A biquad circuit is a typical second-order building block, and it is used for configuring various types of high-order circuits. It is required for circuit designers that the circuit enables some kinds of circuit responses without any component matching constraints, and that it has orthogonal or independent adjusting for characteristic parameters. Additionally, the circuit configuration employed grounded passive components is much recommended for CMOS implementation.

In actuality, active circuit may be claimed to operate in the mixed-modes (i.e. the VM, CM, trans-admittance-mode (TAM) and trans-impedance-mode (TIM)) from a broad viewpoint. So far, there are some research reports concerning OTA-based mixed-mode biquads [1-3]. But the OTA-based biquads above are not applicable to wideband behavior compared with the DVCC-based ones. If the mixed-mode biquad utilizing the PO-DVCCs is devised, the circuit has some excellent performances (i.e. wideband operation, low power dissipation, etc.) in comparison with conventional ones. However, such PO-DVCC-based biquad has not been well researched as yet.

This article focuses on a novel mixed-mode universal biquad using only the PO-DVCCs with grounded passive components as mentioned above. The biquad circuit configuration is consisted of an integrator loop structure with loss-less integrators [4]. The circuit performs the VM, CM, TAM and TIM operations with the selection of the input and output terminals, and it enables the low-pass (LP), band-pass (BP), high-pass (HP), band-stop (BS) and all-pass (AP) characteristics by suitably choosing the input signals with no component matching constraints. The characteristic parameters ω_0 and Q are set orthogonally by the circuit components. Additionally, the sensitivity analysis leads that the

*Dr. Takao Tsukutani, 2682-1 Watari, Sakaiminato, Tottori, 684-0072 Japan, Contact No 18M-06-086 & Email: t.tsukutani@sea.chukai.ne.jp

biquad has extremely low sensitivities with respect to circuit components.

A versatile mixed-mode biquad employing current controlled DVCCs (DVCCCs) is inducted. A DVCCC is an active device utilized effectively the parasitic resistance at the x-terminal of the DVCC. The circuit is composed of plus current output DVCCCs (PO-DVCCCs) and grounded capacitors. And the circuit performance can be controlled electronically with the bias currents of the PO-DVCCCs.

A biquad circuit design is carried out with PSPICE simulation, and the simulation responses are favorable sufficient over a wideband of frequencies. The biquad has several excellent performances in terms of the wideband operation, low power dissipation, orthogonal adjusting for the characteristic parameters, etc., and it suits well for CMOS implementation.

2. Plus current output DVCC

The symbol of the PO-DVCC is given in Figure 1, and hereinto it shows dual plus current output DVCC.

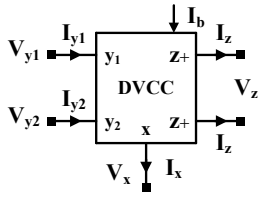


Figure 1: Symbolic representation of PO-DVCC

The PO-DVCC is characterized by the terminal equations [8] below:

$$V_x = V_{y1} - V_{y2}, \quad I_z = I_x \tag{1}$$

The MOS PO-DVCC configuration [11] is shown in Figure 2. It is an active device modified differential difference current conveyor (DDCC) [12], namely one y-terminal with plus polarity is grounded in the DDCC.

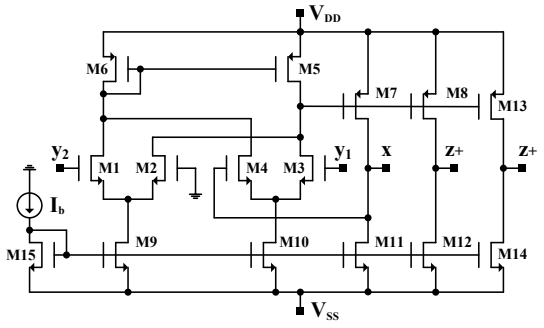


Figure 2: Plus current output DVCC configuration with MOS transistors

3. DVCC-based biquad and its performances

The mixed-mode universal biquad circuit configuration is shown in Figure 3. The circuit is consisted of four single PO-DVCCs, one dual PO-DVCC with grounded resistors and capacitors. This circuit is configured utilizing a second-order loss-less integrator loop structure [4].

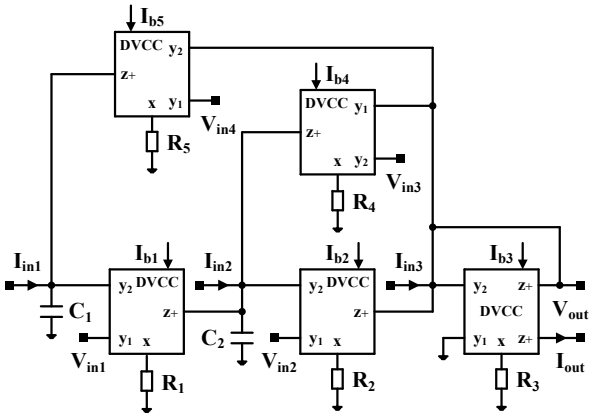


Figure 3: DVCC-based biquad circuit configuration

The voltage and current outputs $V_{out}(s)$, $I_{out}(s)$ are given as:

$$V_{out}(s) = \frac{N_v(s)}{D(s)} \tag{2} \qquad I_{out}(s) = \frac{N_i(s)}{D(s)} \tag{3}$$

where

$$N_v(s) = R_3[\{I_{in3}(s) + \frac{V_{in2}(s)}{R_2}\}s^2 - \frac{1}{C_2R_2}\{I_{in2}(s) + \frac{V_{in1}(s)}{R_1} - \frac{V_{in3}(s)}{R_4}\}s + \frac{1}{C_1C_2R_1R_2}\{I_{in1}(s) + \frac{V_{in4}(s)}{R_5}\}] \tag{4}$$

$$N_i(s) = -\frac{1}{R_3}N_v(s) \tag{5}$$

$$D(s) = s^2 + \frac{R_3}{C_2R_2R_4}s + \frac{R_3}{C_1C_2R_1R_2R_5} \tag{6}$$

It is obvious from the equations above that the circuit performs the mixed-mode behavior with the selection of the input and output terminals, and it enables the various kinds of the circuit responses choosing the input signals suitably.

In the CM operation (i.e. $V_{in1}(s)=V_{in2}(s)=V_{in3}(s)=V_{in4}(s)=0$), the circuit transfer functions are realized as below:

LP: $I_{in1}(s)=I_{in}(s)$, $I_{in2}(s)=I_{in3}(s)=0$

$$T_{LP}(s) = \frac{I_{out}(s)}{I_{in}(s)} = -\frac{1/C_1C_2R_1R_2}{D(s)} \tag{7}$$

BP: $I_{in2}(s)=I_{in}(s)$, $I_{in1}(s)=I_{in3}(s)=0$

$$T_{BP}(s) = \frac{I_{out}(s)}{I_{in}(s)} = \frac{(1/C_2R_2)s}{D(s)} \tag{8}$$

HP: $I_{in3}(s)=I_{in}(s)$, $I_{in1}(s)=I_{in2}(s)=0$

$$T_{HP}(s) = \frac{I_{out}(s)}{I_{in}(s)} = -\frac{s^2}{D(s)} \tag{9}$$

BS: $I_{in1}(s)=I_{in3}(s)=I_{in}(s)$, $I_{in2}(s)=0$

$$T_{BS}(s)=\frac{I_{out}(s)}{I_{in}(s)}=-\frac{s^2+1/C_1C_2R_1R_2}{D(s)} \quad (10)$$

AP: $I_{in1}(s)=I_{in2}(s)=I_{in3}(s)=I_{in}(s)$

$$T_{AP}(s)=\frac{I_{out}(s)}{I_{in}(s)}=-\frac{s^2-(1/C_2R_2)s+1/C_1C_2R_1R_2}{D(s)} \quad (11)$$

The characteristic parameters ω_0 and Q are represented as:

$$\omega_0=\sqrt{\frac{R_3}{C_1C_2R_1R_2R_5}},\quad Q=R_4\sqrt{\frac{C_2R_2}{C_1R_1R_3R_5}} \quad (12)$$

Thus the circuit enables five basic circuit responses with no component matching constraints, and the characteristic parameters ω_0 and Q are set orthogonally by the passive components.

Especially the deviation of the circuit components affects the circuit performance. Hence we examine the effects with sensitivity analysis. The component sensitivities concerning ω_0 and Q are shown in Table 1. These values clarify that the circuit has low sensitive performance. It is an additional note that the sensitivities are not dependent on the component values.

Table 1: Component sensitivities

x	ω_0 sensitivity	Q sensitivity
R_1	-0.5	-0.5
R_2	-0.5	0.5
R_3	0.5	-0.5
R_4	0.0	1.0
R_5	-0.5	-0.5
C_1	-0.5	-0.5
C_2	-0.5	0.5

Successively we examine the performance of the VM circuit. The circuit responses in the VM operation (i.e. $I_{in1}(s)=I_{in2}(s)=I_{in3}(s)=0$) are found with the choice of the input voltages as below:

LP: $V_{in4}(s)=V_{in}(s)$, $V_{in1}(s)=V_{in2}(s)=V_{in3}(s)=0$

$$T_{LP}(s)=\frac{V_{out}(s)}{V_{in}(s)}=\frac{R_3/C_1C_2R_1R_2R_5}{D(s)} \quad (13)$$

BP: $V_{in1}(s)=V_{in}(s)$, $V_{in2}(s)=V_{in3}(s)=V_{in4}(s)=0$

$$T_{BP}(s)=\frac{V_{out}(s)}{V_{in}(s)}=-\frac{(R_3/C_2R_1R_2)s}{D(s)} \quad (14)$$

HP: $V_{in2}(s)=V_{in}(s)$, $V_{in1}(s)=V_{in3}(s)=V_{in4}(s)=0$

$$T_{HP}(s)=\frac{V_{out}(s)}{V_{in}(s)}=\frac{R_3}{R_2}\frac{s^2}{D(s)} \quad (15)$$

BS: $V_{in2}(s)=V_{in4}(s)=V_{in}(s)$, $V_{in1}(s)=V_{in3}(s)=0$

$$T_{BS}(s)=\frac{V_{out}(s)}{V_{in}(s)}=\frac{R_3}{R_2}\frac{s^2+1/C_1C_2R_1R_5}{D(s)} \quad (16)$$

AP: $V_{in1}(s)=V_{in2}(s)=V_{in4}(s)=V_{in}(s)$, $V_{in3}(s)=0$

$$T_{AP}(s)=\frac{V_{out}(s)}{V_{in}(s)}=\frac{R_3}{R_2}\frac{s^2-(1/C_2R_1)s+1/C_1C_2R_1R_5}{D(s)} \quad (17)$$

The VM circuit can also realize the LP, BP, HP, BS and AP transfer functions like as the CM one. The characteristic parameters ω_0, Q and component sensitivities are same as the current-mode ones.

In addition, the TAM and TIM operations are performed selecting the input terminal $I_{in}(s)/V_{in}(s)$ and output terminal $V_{out}(s)/I_{out}(s)$, respectively. Also, the circuit responses are derived with the above-mentioned ways in these operational modes.

4. Current controlled DVCC and biquad circuit configuration

In reality, the DVCC has parasitic resistance existing at the x-terminal, and the resistance varies with the bias current. The parasitic resistance R_x is given as:

$$R_x=K\left(\mu C_{ox}\frac{W}{L}I_b\right)^{-\frac{1}{2}} \quad (18)$$

where μ, C_{ox}, W/L, I_b and K denote the electron mobility, gate oxide capacitance per unit area, MOS transistor aspect ratio, bias current and constant parameter, respectively. Equation (18) shows that the parasitic resistance R_x is electronically adjusted with the bias current I_b. An active device made good use of the parasitic resistance is named DVCCC.

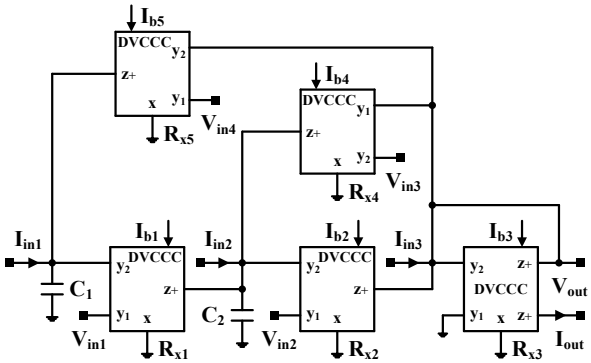


Figure 4: DVCCC-based biquad circuit configuration

Figure 4 illustrates the biquad circuit configuration employing the PO-DVCCCs. The circuit is composed of the PO-DVCCCs and grounded capacitors. The voltage and current outputs $V_{out}(s)$, $I_{out}(s)$ are given replacing the external resistance R_i to the parasitic resistance R_{xi} from (4) to (17). The characteristic parameters ω_0 and Q are also obtained with the same replacement.

The PO-DVCCC-based biquad has electronic tunability for the characteristic parameters with the bias currents. In addition, it can

be expected that the circuit has a high potential for low power performance, because the circuit configuration is without external resistors.

5. A circuit design and simulation responses

In order to confirm our proposal, we tried to realize a CM circuit using PSPICE simulation. As a circuit design, we deal a specification with the cut-off frequency f_0=1MHz, quality factor Q=1.0 and gain constant H=1.0. A macro model of the PO-DVCC in Figure 2 was used in this simulation.

To achieve the specification above, we set that the circuit components were R_i (i=1, 2, 3, 4, 5)=12kΩ and C_1=C_2=12pF, and that the bias currents (i=1, 2, 3, 4, 5)=10μA, DC supply voltages V_{DD}=-V_{SS}=0.8V and input current I_{in}=10μA, respectively.

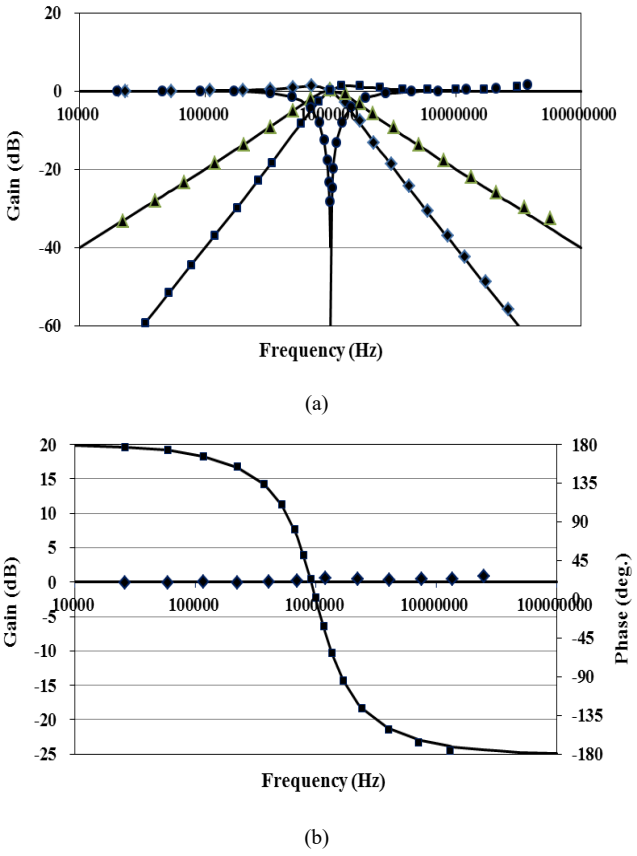


Figure 5: Simulated circuit responses

The simulation responses are given in Figure 5. Figure 5 (a) presents the LP, BP, HP and BS responses, and meanwhile the AP response is indicated in Figure 5 (b). In the figures, the marks signify the simulated responses, and the continuous lines show the theoretical responses. The simulated responses agree very well with the ideal values over a wide range of frequencies. The consumed power dissipation was 0.439mW. It's rather small.

The MOS transistor aspect ratios [10] are given in Table 2. Additionally, we utilized device parameters of MOSIS 0.5μm for other parameters.

Table 2: MOS transistor aspect ratios.

MOS transistors	W / L
M1-M4	20μm / 0.5μm
M5-M8, M13-M15	30μm / 2μm
M9-M12, M16-M19	10μm / 2μm

6. Conclusions

In this paper, we have proposed a novel mixed-mode universal biquad employing plus current output DVCCs with grounded passive components. Additionally, we demonstrated that the circuit accomplishes the mixed-mode behavior, and it enables five kinds of basic circuit responses without any component matching constraints. The circuit is capable of orthogonally adjusting the characteristic parameters by the circuit components. It has also been revealed that the component sensitivities were extremely low in the circuit.

The versatile plus current output DVCCC-based biquad has been inducted. The circuit can adjust electronically the circuit performances with the bias currents.

A circuit design was performed with SPICE simulation, and the simulated responses were sufficient good over a wideband of frequencies.

Our proposed biquad possesses several advantages concerning the wideband operation, low power dissipation, orthogonal adjusting for the characteristic parameters, etc., and it is easily capable of integrating in CMOS technology.

In reality, the DVCC has non-ideal elements (i.e. voltage and current tracking errors, etc.), and especially they affect the circuit performances. An appropriate solution on this must be found out in the future.

Acknowledgment

The authors would like to thank to Mr. David Martin for his helpful suggestions. Especially we are grateful to him for editing our manuscript properly.

References

[1] M.T. Abuelma'atti, et al, "A novel mixed-mode OTA-C universal filter" Int. J. Electron., 92(7), 375-383, 2005. https://doi.org/10.1080/08827510412331295009

[2] C.N. Lee, "Multiple-mode OTA-C universal biquad filters" Cir. Syst. Sig. Proc., 29, 263-274, 2010. DOI 10.1007/s00034-009-9145-0

[3] T. Tsukutani, et al, "A mixed-mode biquad employing OTAs and grounded capacitors" J. Elec. Eng., 6(3), 151-155, 2018. DOI: 10.17265/2328-2223/2018.03.003

[4] E. Sanchez-Sinencio, et al, "Generation of continuous-time two integrator loop OTA filter structures" IEEE Trans. Cir. Syst., 35(8), 936-946, 1988. DOI: 10.1109/31.1840

[5] Y. Tao, J.K. Fidler, "Electronically tunable dual-OTA second-order sinusoidal oscillators/filters with non-interacting controls: a systematic synthesis approach" ibid., 47(2), 117-129, 2000. DOI: 10.1109/81.828566

[6] S.H. Tu, et al, "Novel versatile insensitive universal current-mode biquad employing two second-generation current-conveyors" Int. J. Electron., 89(12), 897-903, 2002. https://doi.org/10.1080/0020721031000121262

[7] J.W. Horng, "High input impedance voltage-mode universal biquadratic filters with three in-puts using plus-type CCIIs" ibid., 91(8), 465-475, 2004. https://doi.org/10.1080/00207210412331294612

[8] H.O. Elwan, A.M. Soliman, "Novel CMOS differential voltage current conveyor and its applications" IEE Proc. Cir. Dev. Syst., 144(3), 195-200, 1997. DOI: 10.1049/ip-cds:19971081

[9] M.A. Ibrahim, et al, "A 22.5MHz current-mode KHN-biquad using differential voltage current conveyor and grounded passive elements" Int. J. Electron. Com., 59, 311-318, 2005. https://doi.org/10.1016/j.aeue.2004.11.027

[10] S. Minaei, M.A. Ibrahim, "General configuration for realizing current-mode first-order all-pass filter using DVCC" Int. J. Electron., 92(6), 347-356, 2005. https://doi.org/10.1080/00207210412331334798

[11] T. Tsukutani, et al, "Novel current-mode bi-quadratic circuit using only plus type DO-DVCCs and grounded passive components" ibid., 94(12), 1137-1146, 2007. DOI: 10.1080/00207210701791085

[12] W. Chiu, et al, "CMOS differential difference current conveyors and their applications" IEE Proc. Cir. Dev. Syst., 143(2), 91-96, 1996. DOI: 10.1049/ip-cds:19960223

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 238-251 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

An Overview of Data Center Metrics and a Novel Approach for a New Family of Metrics

Moises Levy*, Daniel Raviv

Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, 33431, FL, USA

ARTICLE INFO

Article history:
Received: 13 February, 2018
Accepted: 04 April, 2018
Online: 11 April, 2018

Keywords:
Data center
Mission critical
Metric
Measurement
Performance
Efficiency
Productivity
Sustainability
Operations
Infrastructure
Management
Risk
Site

ABSTRACT

Data centers' mission critical nature, significant power consumption, and increasing reliance on them for digital information, have created an urgent need to monitor and adequately manage these facilities. Metrics are a key part of this effort as their indicators raise flags that lead to optimization of resource utilization. A thorough review of existing data center metrics presented in this paper shows that while existing metrics are valuable, they overlook important aspects. New metrics should enable a holistic understanding of the data center behavior. This paper proposes a novel framework using a multidimensional approach for a new family of data center metrics. Performance is examined across four different sub-dimensions: productivity, efficiency, sustainability, and operations. Risk associated with each of those sub-dimensions is contemplated. External risks are introduced, namely site risk, as another dimension of the metrics, and makes reference to a methodology that explains how it is calculated. Results from metrics across all sub-dimensions can be normalized to the same scale and incorporated in one graph, which simplifies visualization and reporting. The new family of data center metrics can help to standardize a process that evolves into a best practice to help evaluate data centers, to compare them to each other, and to improve the decision-making process.

1. Introduction

The ongoing significant and increasing reliance on digital information has led data centers to play a key role in guaranteeing that information is constantly available for users, and adequately stored. A novel framework for data center metrics using a multidimensional approach was introduced in a paper originally presented at the 15th LACCEI International Multi-Conference for Engineering, Education, and Technology: Global Partnerships for Development and Engineering Education in 2017 [1], for which this work is an extension.

Data centers are energy intensive complexes, and this sector is expected to grow substantially. These circumstances have prompted the desire and need to make data centers more efficient and sustainable, while at the same time ensuring reliability and availability. Efforts undertaken to pursue this goal include new legislation, the development of standards and best practices to follow when designing, building and operating a data center, and metrics to monitor performance and find areas of improvement.

*Corresponding Author: Moises Levy. Florida Atlantic University, USA
Email: mlevy2015@fau.edu

Being that data centers are a growing field subject to new and evolving technologies, current practices ignore important pieces of information. Existing data center metrics are very specific, and fail to take into consideration the holistic performance of the data center. In addition, there is an imminent need for metrics to incorporate an assessment of the risk to which the data center is exposed.

The proposed concept addresses concerns from the recent United States Data Center Usage Report (June 2016) [2], which communicates the need to expand research on data center performance metrics that better capture efficiency, in order to identify and understand areas of improvement. The main motivation of this paper is to consolidate existing metrics and current practices, explain areas of improvement, and finally propose a novel multidimensional approach for data center metrics incorporating productivity, efficiency, sustainability, and operations, as well as measurements of all the different risks associated to the data center. This paper adds technical and scientific support to existing theoretical and practical work that has mostly been carried out from outside academia. The ultimate goal

of this work is to help standardize a process that eventually becomes a best practice to rate data centers.

2. Background

A data center can be defined as a dedicated facility with all the resources required for storage, processing and sharing digital information, and its support areas. Data centers comprise the required infrastructure (e.g., power distribution, environmental control systems, telecommunications, security, fire protection, and automation) and information technology (IT) equipment (including servers, storage and network/communication equipment). Data centers are very dynamic; equipment can be upgraded frequently, new equipment may be added, obsolete equipment may be removed, and old and new systems may be in use simultaneously. Several threats can cause failures in a data center, including technical issues and human errors. The cost of downtime depends on the industry, and could reach thousands of dollars per minute. Recent reports show that the average cost associated with data center downtime for an unplanned outage is approximately $ 9,000 per minute, an increase of about 60% from 2010 to 2016 [3].

According to standards, best practices, and user requirements, the infrastructure of data centers must comply with stringent technical requirements that guarantee reliability, availability, and security, as they highly correlate with cost and efficiency. An infrastructure with high reliability and availability must have system redundancy, which makes it more expensive, and probably less efficient [4]. In this context, reliability is the ability of the system to perform its functions under stated conditions for a specified period of time, whereas availability refers to the degree to which a system is operational when it is required for use. Redundancy is the multiplication of data center components used to enhance reliability, since they may malfunction at a certain point due to maintenance, upgrade, replacement, or failure [5].

Data centers can consume 40 times more energy than conventional office buildings. IT equipment by itself can consume 1100 W/m^2 [6], and its high concentration in a data center results in higher power densities, compared to conventional office buildings, where energy consumption ranges between 30 and 110 W/m^2. In addition, the energy consumption profile of data centers is very different from conventional office buildings, since IT equipment power consumption represents more than 50% of the total consumption in data centers. Figure 1 shows an example of energy consumption breakdown [7].

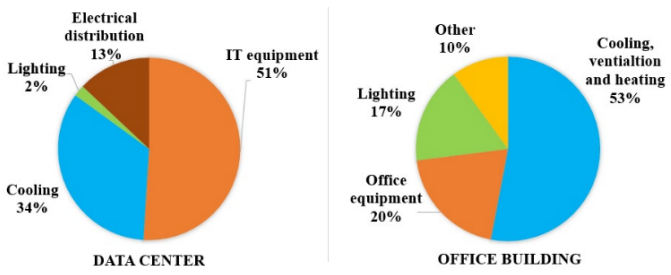


Figure 1: Energy consumption profile in a data center and office building

The United States is home to approximately 3 million data centers, representing one data center for every 100 people [8]. Data center electricity consumption increased approximately 24% from 2005-2010, 4% from 2010-2014, and is expected to grow 4% from 2014-2020 [2]. In 2014, data centers in the United States consumed around 70 billion kWh, which is 1.8% of total electricity consumption, and are predicted to consume around 73 billion kWh in 2020 [2]. Emerging technologies and energy management strategies may decrease the projected energy consumption.

Environmental impact of data centers varies depending on the energy sources used and the total heat generated. The differences between the lowest and the highest greenhouse gas (GHG) emissions associated with each energy source is approximately a factor of 200 [9]. For the period from 2002 to 2020, the emissions associated with the IT sector and the data center sector are estimated to grow by 180% and 240% respectively, considering business as usual [10]. This growth rate is much faster compared to the 30% increase in total emissions from all sources, again considering business as usual. Conversely, the IT sector has contributed to a reduction in emissions in other sectors, since IT is an enabling infrastructure for the global economy. For 1 kWh consumed by the IT sector in the United States, other 10 kWh are saved in other sectors due to the increase in economic productivity and energy efficiency [11]. The growing ubiquity of IT driven technologies has revolutionized and optimized the relation between efficiency and productivity, and energy consumption across every sector of the economy.

The deregulation of telecommunications in the United States through the Telecommunications Act of 1996 promoted the creation of a number of standards and best practices related to telecommunications, and more recently to data centers. Standards provide the most complete and reliable guidance to design or assess a data center, from national/ local codes (required) to performance standards (optional). Due to their mission critical tasks and elevated costs, data centers must be designed, built and operated in compliance with those standards to ensure basic performance and efficiency. Standards, however, do not translate into best practices when the objective is attaining the highest possible reliability and availability, under the best performance. Optional standards and best practices have contributed to achieving this goal.

Data centers standards and best practices evolve continually adapting to emerging needs, and addressing new issues and challenges. The *ASHRAE Technical Committee 9.9* has created a set of guidelines regarding the optimal and allowable range of temperature and humidity set points for data centers [12] [13] [14]. The *ANSI/ASHRAE standard 90.4 (Energy Standard for Data Centers)* establishes the minimum threshold for data center energy efficient design, construction, operation maintenance, and utilization of renewable energy resources [15]. The *Singapore Standard SS 564 (Green Data Centers)* addresses planning, building, operation and metrics of green data centers [16]. Standards contribute to classify data centers based on their reliability and infrastructure redundancy, such as the *ANSI/BICSI-002* (class F0 to F4) [5], the *ANSI/TIA-942* (rating 1 to 4) [17] and the *Data Center Site Infrastructure Standard* from Uptime Institute (tier levels 1 to 4) [18] [19].

Likewise, there is significant concern among legislators about how efficiently an IT facility uses energy. Governments have also imposed regulations on data centers depending on the nature of the business The *Energy Efficiency Improvement Act of 2014* (H.R. 2126) demands federal data centers to implement energy efficiency standards. This is motivated by the fact that federal data centers energy consumption represents 10% of all data centers in the United States. This bill encourages federal data centers to improve energy efficiency and develop best practices to reduce energy consumption [20]. According to the *Data Center Optimization Initiative* (DCOI), federal data centers must reduce their power usage efficiency below a specified threshold, unless they are scheduled to be shutdown, as part of the *Federal Data Center Consolidation Initiative* (FDCCI). In addition, federal data centers must replace manual collection and reporting with automated infrastructure management tools by the end of 2018, and must address different metric targets including energy metering, power usage effectiveness, virtualization, server utilization and facility monitoring [21].

Organizations that have contributed to the creation of standards, white papers, best practices and other documents related to the Data Center industry are: the International Organization for Standardization (ISO), the National Institute of Standards and Technology (NIST), the Institute of Electrical and Electronics Engineers (IEEE), the International Electrotechnical Commission (IEC), the National Fire Protection Association (NFPA), the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE), the U.S. Department of Energy (DOE), the U.S. Environmental Protection Agency (EPA), the U.S. Green Building Council (USBGC), the Telecommunication Industry Association (TIA), the Building Industry Consulting Service International (BICSI), the Uptime Institute, The Green Grid (TGG), the Association for Computer Operations Management (AFCOM), the International Computer Room Expert Association (ICREA), the International Data Center Authority (IDCA), the European Commission (EU Code of Conduct for Data Centres), the British Computer Society (BCS), the Japan's Green IT Promotion Council (GIPC), the Japan Data Center Council (JDCC), and the Singapore Standards Council among others.

3. Overview of Data Center Metrics

Metrics are measures of quantitative assessment that allow comparisons or tracking of performance, efficiency, productivity, progress or other parameters over time. Through different metrics data centers can be evaluated in comparison to goals established, or to similar data centers. Variations or inconsistencies in measurements can produce a false result for a metric, which is why it is very important to standardize metrics. Much of the current metrics, standards and legislation on data centers is focused towards energy efficiency, as this has proven to be a challenge given the rapid growth of the sector and its energy intensive nature.

One of the most widely used energy efficiency metrics is *Power Usage Effectiveness* (PUE), introduced in 2007 [22]. Figure 2 shows the energy flow in a data center. PUE is calculated as the ratio of energy used in a facility to energy delivered to IT equipment [23]. Similarly, *Data Center infrastructure Efficiency* (DCiE) is defined as the reciprocal of PUE [24] with a value ranging between 0 and 1 to make the metric easier to understand in terms of efficiency. Organizations such as the EPA have selected PUE as the metric to analyze energy performance in data centers, and define Source PUE as the ratio of total facility energy used to UPS energy [25], [26]. The main limitation of PUE is that it only measures the efficiency of the building infrastructure supporting a given data center, but it indicates nothing about the efficiency of IT equipment, or operational efficiency, or risk involved [2] [27] [28].

Multiple metrics have been developed to measure other aspects of efficiency. *Data Center Size metric* (e.g., 'mini', 'small', 'medium', 'large', 'massive', and 'mega' data centers) is based on the number of racks and the physical compute area of the data center [29]. Other classifications based on the size of the data center have been proposed such as 'hyperscale', 'service provider', 'internal', 'server room', and 'server closet' [2]. The rack *Density metric* (e.g., low, medium, extreme, and high) considers the measured peak power consumption on every rack (density for rack) and across the compute space [29]. *Data Center Density* (DCD) is defined as the total power consumption of all equipment divided by the area. *Compute Power Efficiency* (CPE) is estimated as the IT equipment utilization multiplied by the IT equipment power consumption and divided by the total facility power consumption [30].

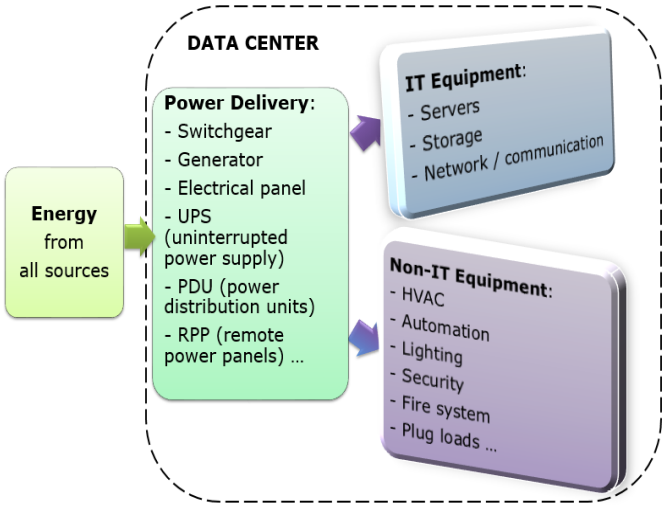


Figure 2: Data center energy flow diagram

Metrics have been proposed to measure server efficiency and performance for computer servers and storage. The metric *FLOPS* (floating-point operations per second) *per Watt* measures performance per unit of power [31], and is used to rank the most energy efficient supercomputers' speed. Standard Performance Evaluation Corporation (SPEC) has contributed with proposals such as *SPEC Power and Performance Benchmark Methodology*, which are techniques recommended for integrating performance and energy measures in a single benchmark; *SPECpower_ssj2008*, a general-purpose computer server energy-efficiency measure, or power versus utilization; *SPECvirt_sc2013* for consolidation and virtualization; *SPEComp2012* for highly parallel complex computer calculations; *SPECweb2009* for web applications [32]. The Transaction Processing Performance Council (TPC) uses benchmarks with workloads that are specific to database and data

management such as: *TPC-Energy*, which focuses on energy benchmarks (e.g. *TPC-C* and *TPC-E* for online transaction processing, *TPC-H* and *TPC-DS* for business intelligence or data warehouse applications, *TPC-VMS* for virtualized environment) [32]. VMWare proposed *VMmark* to measure energy consumption, performance and scalability of virtualization platforms [32]. The Storage Networking Industry Association (SNIA) Emerald Program and the Storage Performance Council (SPC) have contributed to establishing measurements for storage performance and the power consumption associated with the workloads [32]. The *Space Wattage and Performance* metric (SWaP = Performance/ (Space x Power)) [33] incorporates the height in rack units (space), the power consumption (measured during actual benchmark runs or taken from technical documentation), and the performance (measured by industry standard benchmarks such as SPEC).

For cooling and ventilation system efficiency, metrics have also been proposed. *HVAC effectiveness* is measured through the ratio of IT equipment energy consumption to HVAC system energy consumption. *Airflow Efficiency* shows the total fan power needed by unit of airflow (total fan power in Watts / total fan airflow in cfm). *Cooling System Efficiency* is calculated as the ratio of average cooling system power consumption divided by the average data center cooling load [26]. *Cooling System Sizing Factor* shows the ratio between the installed cooling capacity and the peak cooling load. *Air Economizer Utilization Factor* and *Water Economizer Utilization Factor* measure usage at full capacity over a year in percentage terms. *Air temperature* metric measures the difference between the supply and return air temperature in the data center. *Relative humidity* metric measures the difference between the return and supply air humidity in the data center [2]. The *Rack Cooling Index* (RCI) gauges cooling efficiency for IT equipment cabinets compared to the IT equipment intake temperature [34]. The *Return Temperature Index* (RTI) gauges the performance of air management systems [35].

The *Coefficient of Performance of the Ensemble* (COP) gauges the ratio of the total heat load to power consumption of the cooling system [36]. The Cooling Capacity Factor (CCF) estimates the utilization of the cooling capacity, by dividing total rated cooling capacity by the UPS output multiplied by 110% [37]. The *Energy Efficiency Ratio* (EER) is the ratio of the cooling capacity to power input at 95 ºF, and the *Seasonal Energy Efficiency Ratio* (EER) is the ratio of the total heat removed during the annual cooling season by total energy consumed in the same season [38]. The *Sensible Coefficient of Performance* (SCOP) is defined as the ratio of net sensible cooling capacity divided by total power required to produce that cooling (excluding reheat and humidifier) n consistent units [39] [40]. This metric was chosen to computer room air conditioner due to the unique nature and operation of data center facilities.

For electrical equipment different metrics have been proposed [2]. *UPS load factor* gauges the relation between the peak value and the nominal capacity. *UPS system efficiency* shows the relation of the output power and the input power. *Lighting density* shows the lighting power consumption per area.

Data center sustainability metrics have also been introduced. The *Green Energy Coefficient* (GEC) measures the percentage of total energy sourced from alternative energy sources, such as solar, wind, or geothermal plants, and encourages the use of renewable energy [41]. *Carbon Usage Effectiveness* (CUE) measures total CO_2 emissions in relation to IT equipment energy consumption [42]. *Water Usage Effectiveness* (WUE) measures total water usage in relation to IT equipment energy consumption [43]. *Energy Reuse Effectiveness* (ERE) gauges how energy is reused outside of the data center. It is calculated as the total energy minus reuse energy divided by IT equipment energy, or the *Energy Reuse Factor* (ERF) calculated as the reuse energy divided by the total energy [44]. The *Electronic Disposal Efficiency* (EDE) measures how responsible the discarded electronic and electrical equipment are managed, as the ratio of equipment disposed of through known responsible entities and the total of equipment disposed [45].

Other metrics quantify energy consumption related to environmental sustainability. The following categories of metrics are defined: IT strategy, IT hardware asset utilization, IT energy and power efficient hardware deployment, and site physical infrastructure overhead [46]. For those metrics different factors are defined, such as, the *Site-Infrastructure Power Overhead Multiplier* (SI-POM) estimated as the power consumption at the utility meter divided by the total power consumption at the plug of all IT equipment; the *IT Hardware Power Overhead Multiplier* (H-POM), estimated as the ratio of the AC hardware load at the plug and the DC hardware compute load, showing the IT equipment efficiency; the *Deployed Hardware Utilization Ratio* (DH-UR), estimated as the number of servers running live applications divided by the total servers deployed, or as the ratio of terabytes of storage holding data and the total of terabytes of storage deployed; the *Deployed Hardware Utilization Efficiency* (DH-UE) measured as the ratio of minimum number of servers required for peak compute load and the total number of servers deployed, which shows the possibilities of virtualization; and free cooling [46]. The *Free Cooling* metric estimates potential savings using outside air; and the *Energy Save* metric calculates the amount of money, energy, or carbon emission savings that accrue if IT equipment hibernates while it is not in use [46].

After recognizing the need for performance metrics that better capture the efficiency of a given data center, different entities have proposed metrics that measure the functionality of the data center (e.g. amount of computations it performs) and relate that to energy utilization. An example of these are metrics to track useful work produced at a data center compared to power or energy consumed producing it. “Useful work” is defined as the tasks executed in a period of time, each one having a specific weight related to its importance.

Different productivity metrics have also been proposed. Although they are often specific to each user’s activity, these metrics provide a framework for comparison. The *Data Center Performance Efficiency* metric is the ratio of useful work to total facility power [30]. The *Data Center Energy Productivity* (DCeP) metric is defined as the amount of useful work produced, divided by the total data center energy consumed producing it [47]; The *Data Center Compute Efficiency* (DCcE) metric gauges the efficiency of compute resources, intended to find areas of

improvement [48]; The *Data Center Storage Productivity* (DCsP) metric expresses the ratio of useful storage system work to energy consumed [49].

Fixed and proportional overhead metrics help analysts and managers understand how energy and cost influences the use of IT equipment. The fixed portion of energy consumption considers use when all IT equipment are unused. The variable part takes into account IT equipment load [50]. The *Data Center Fixed to Variable Energy Ratio* (DC-FVER) metric, defined as the fixed energy divided by the variable energy plus one (1 + fixed energy / variable energy). It shows the inefficiencies through the wasted energy not delivering 'useful work'. The metric reflects the proportion of energy consumption that is variable, considering IT equipment, software and infrastructure [51]. Metrics related to energy-proportional computing, based on computing systems consuming energy in proportion to the work performed have been proposed. The *Idle-to-Peak power Ratio* (IPR) metric is defined as the ratio system's idle consumption with no utilization over the full utilization power consumption [52]. The *Linear Deviation Ratio* (LDR) metric, shows how linear power is, compared to utilization curve [52].

Other metrics have been also proposed. The *Digital Service Efficiency* (DSE) metric shows the productivity and efficiency of the infrastructure through performance, cost, environmental impact, and revenue. Performance is measured by transactions (buy or sell) per energy, per user, per server and per time. Cost is measured by amount of money per energy, per transaction and per server. Environmental impact is estimated in metric tons of carbon dioxide per energy and per transaction. Revenue is estimated per transaction and per user [53]. The *Availability, Capacity and Efficiency* (ACE) performance assessment factors in the availability of IT equipment during failures, the physical capacity available, and how efficient the cool air delivery to IT equipment is [54].

Metrics that combine measurements of efficiency and productivity have been also proposed. The *Corporate Average Datacenter Efficiency* (CADE), estimated as the IT equipment efficiency factored by the facility efficiency (CADE = IT equipment efficiency × facility efficiency = IT equipment asset utilization × IT equipment energy efficiency × Site asset utilization × Site energy utilization) [55]. The *Data Center Energy Efficiency and Productivity* (DC-EEP) index results from multiplying the *IT Productivity per Embedded Watt* (IT-PEW) and the *Site Infrastructure Energy Efficiency ratio* (SI-EER) [56].The IT organization is mainly responsible for the IT-PEW. The SI-EER is measured dividing the power required for the whole data center by the conditioned power delivered to the IT equipment.

The *Datacenter Performance per Energy* (DPPE) considers four sub-metrics: *IT Equipment Utilization* (ITEU), *IT Equipment Energy Efficiency* (ITEE), *Power Usage Effectiveness* (PUE) and *Green Energy Coefficient* (GEC). ITEU is the ratio of total measured power of IT equipment to total rated power of IT equipment. ITEE is the ratio of total rated capacity of IT equipment and total rated energy consumption of IT equipment. PUE is the total energy consumption of the data center divided by the total energy consumption of IT equipment, which promotes energy saving of facilities. GEC is the green energy divided by the total energy consumption of the data center, which promotes the use of green energy. Then the metric is defined as DPPE = ITEU × ITEE × 1/PUE × 1/ (1-GEC) [57].

The *Performance Indicator* (PI) metric was introduced to visualize the data center cooling performance in terms of the balance of the following metrics: thermal conformance, thermal resilience and energy efficiency [58] [59]. IT thermal conformance indicates the proportion of IT equipment operating within recommended inlet air temperatures ranges during normal operation. IT thermal resilience shows if there is any equipment at risk of overheating in case redundant cooling units are not operating due to a failure or planned maintenance. Energy efficiency is measured through the PUE ratio, and it indicates how the facility is operated compared to pre-established energy efficiency ratings.

Different measurements for operational performance have been proposed recently. The methodology *Engineering Operations Ratio* [60] shows the systems operational performance related to its design, as the operational effectiveness for each component (e.g., designed PUE related to actual PUE). The *Data Center Performance Index* [61], takes into account three categories: availability, efficiency and environmental. Availability is measured through the number of incidents or the time of loss of service; performance is gauged through *Power Usage Effectiveness* (PUE) and *Water Usage Effectiveness* (WUE); and environmental is assessed using the Greenhouse Gas (GHG) emissions. Possible indexes are A, B, C, D or not qualifying.

In addition, there have been different proposals to incorporate probability and risk to data center key indicators. The *Class* metric indicates the probability of failure of a data center in the next 12 months, and the associated risk is the Class multiplied by the consequences [62]. It is based on the standard IEEE-3006.7 [63], where *Class* (also called unreliability) is defined as one minus reliability. The *Data Center Risk Index* ranks different countries related to the probability of the factors that affect operations of the facility. Those factors include energy (cost and security), telecommunications (bandwidth), sustainability (alternative energy), water availability, natural disasters, ease of doing business, taxes, political stability, and GDP per capita [64]. The index is mainly designed to contribute on decisions based on the risk profile of the country, although it does not take into account specific business requirements or the fact that some risks can be mitigated.

The authors of this paper have previously proposed a *Data Center Site Risk* metric [65], to help evaluate data center sites and compare them to each other, or to compare different scenarios where the data center operates. The methodology for the *Data Center Site Risk* metric of a specific location is simplified in four steps: the first one is to identify threats and vulnerabilities; the second is to quantify the probability of occurrence of the events; the third one is to estimate potential consequences or impact of the events; and the last one to calculate the total risk level considering weights.

Existing data center metrics reviewed in this paper are listed in Table A.1 (Appendix A), classified by type and main promoter.

Each metric listed uses its own definition for terms such as efficiency, productivity, performance, risk, among others, which must be taken into account when doing comparisons. In addition, there has been academic research in specific data center metrics and risk, such as a modified PUE metric using power demand [66], PUE for a CCHP Natural gas or Biogas Fuelled Architecture [67], PUE for application layers [68], performance metrics for communication systems [69], load dependent energy efficiency metrics [70], workload power efficiency metric [71], airflow and temperature risk [72], power distribution systems risk [73], quality of service and resource provisioning in cloud computing [74][75][76], life cycle cost using a risk analysis model [77], risk for cloud data center overbooking [78], risk management for virtual machines consolidation [79], risk management under smart grid environment [80], and risk for data center operations in deregulated electricity markets [81].

These joint efforts have significantly improved efficiency on the data center infrastructure so that energy consumption has started to flatten out over time [2].

4. Multidimensional Approach for New Family of Data Center Metrics

Existing metrics fail to incorporate important aspects such as the risk involved in processes and operations, for a holistic understanding of the data center behavior. This being the case, comparisons between data center scores with the purpose of evaluating areas of improvement is not an easy task. Furthermore, currently there is no metric that examines performance and risk simultaneously. A data center may have high performance indicators, with a high risk of failure. Research must therefore be refocused to incorporate risks, management and performance. Having this information may work as an early warning system so that mitigation strategies and actions are undertaken on such mission critical facilities.

A new family of metrics can help understand the performance of new and existing data centers, including their associated risk. The proposed novel data center multidimensional scorecard effectively combines performance and risk. Performance is inspected across four different sub-dimensions: productivity, efficiency, sustainability and operations. Risk associated with each of those sub-dimensions is contemplated. External risks are also considered independently of performance, namely site risk.

Figure 3 shows a diagram of the proposed data center multidimensional metric. It is important to highlight that correlation can exist between the different elements of the scorecard; however, for the sake of the explanation this proposal has been simplified by assuming there is no correlation between different performance sub-dimensions. Measurements through different mechanisms are explained below.

4.1 Productivity

Productivity gives a sense of work accomplished and can be estimated through different indicators, such as the ratio of useful work completed to energy usage, or useful work completed to the cost of the data center.

Useful work can be understood as the sum of weighted tasks carried out in a period, such as transactions, amount of information processed, or units of production. The weight of each task is allocated depending on its importance. A normalization factor should be considered to allow the addition of different tasks.



Figure 3: Data center multidimensional metric

Table 1 presents general concepts for these productivity measurements.

Table 1. Productivity measurements

Productivity	Concept
Useful work	- Sum of weighted tasks carried out in a period. - Useful work per energy consumption. - Useful work per physical space. - Useful work to cost of the data center.
Downtime	- Actual downtime, in terms of length, frequency, and recovery time. - A separate measurement within this category will calculate the impact of downtime on productivity, measured as the 'useful work' that was not carried out as well as other indirect tangible and intangible costs due to this failure. To obtain the data a process needs to be established where downtime data (date, time and duration) is sent to this system. To calculate the impact on productivity a scale could be defined based on previous reports of data center outages costs [3].
Quality of service	Quality of service measurements compared to pre-established values. These measurements can include variables in the time domain (e.g., maximum or average waiting time, congestion detection, latency), or scheduling and availability of resources.

Obtaining data to estimate these metrics requires a process for measuring 'useful work', and costs for the specific data center. Cost includes capital expenditures (fixed assets and infrastructure equipment) and operating expenses (energy, human resources, maintenance, insurance, taxes, among other). When including monetary values, and comparing these metrics across time, all future values of money need to be brought to present value so the comparison is consistent. Once processes are in place, calculations of these metrics can be performed in real-time automatically.

4.2 Efficiency

Efficiency has been given substantial attention due to the high

energy consumption of the data center sector. Many initiatives have emerged to measure efficiency. Key indicators show how energy efficient site infrastructure, IT equipment, environmental control systems, and other systems are. Power consumption and utilization data can be directly collected from various equipment elements. It can be measured through different energy efficiency measurements briefly explained in Table 2.

Table 2. Efficiency measurements

Efficiency	Concept
Site infrastructure	The ratio of the energy delivered to IT equipment total to total energy used by the data center The value becomes higher if is more efficient. Promotes energy management in facility.
IT equipment utilization	The ratio of total measured power of IT equipment to total rated power of IT equipment. The value becomes higher if is more efficient. The lowest value if all IT equipment are unused. Promotes efficient operation of IT equipment.
IT equipment efficiency	The ratio of the total potential capacity of IT equipment and the total energy consumption of IT equipment. The value becomes higher with more efficient IT equipment. Promotes the procurement of efficient IT equipment, with higher processing capacity per energy.
Physical space utilization	Physical space used divided by total physical space. Energy consumption of all equipment divided by total physical space. It can also be measured in each cabinet (rack unit or area). Promotes efficient planning of physical space.

4.3 Sustainability

Sustainability can be defined as development that addresses current needs without jeopardizing future generations' capabilities to satisfy their own needs [82].

The sustainability of a data center can be measured in different ways, such as calculating the ratio of green energy sources to total energy, estimating the carbon footprint, or the water usage. In addition, an evaluation may be conducted on how environmentally friendly the associated processes, materials, and components are. Table 3 presents general concepts for these sustainability measurements.

Table 3. Sustainability measurements

Sustainability	Concept
Carbon footprint	Each energy source can be assigned a different value of carbon footprint. Use of existing and widely known methods to evaluate the greenhouse gas emissions (e.g., carbon dioxide, methane, nitrous oxide, fluorinated gases). This estimation can be automated.
Green energy sources	The ratio of green energy consumption to total energy consumption. The data can be obtained automatically from real-time measurements.
Water usage	The ratio of total water used to energy consumption of IT equipment. The data can be obtained automatically from real-time measurements.
Environmentally friendly	How environmentally friendly processes, materials and components are. The information is collected by conducting an analysis or audit of processes. It must be updated if a process changes.

4.4 Operations

Operations measurements gauge how well managed a data center is. This must include an analysis of operations, including site infrastructure, IT equipment, maintenance, human resources training, and security systems, among other factors. Audits of systems and processes are necessary to gather the required data. This data should include factors such as documentation, planning, human resources activities and training, status and quality of maintenance, service level agreement, and security. Table 4 shows general concepts for the operations measurements recommended.

Table 4. Operations measurements

Operations	Concept
Documentation	Procedures and policies for data center should be formally documented. All information should be available in digital format.
Planning	Effective planning is desired to reduce downtime. Planning for maintenance, new components, relocations, upgrades, replacements, and life cycle evaluations are needed.
Organization and human resources.	Organizations with an integrated approach are desired, including interactions between different departments and reporting chain. Personnel with the required qualifications and technical training is needed to properly operate the facility.
Maintenance	Preventive, reactive and deferred maintenance programs are required. Consider the equipment manufacturer or vendor recommendations. Predictive maintenance and failure analysis programs should be included. Maintenance management systems are desired to track the status, frequency and quality of the related activities.
Service level agreement	The commitment that prevails with the service provider, including the quality, availability and responsibilities for the service.
Security	Electronic and physical security. Evaluated and assessed against pre-established scales.

Different organizations have recognized the importance to standardize the operation and management of data centers, and have contributed to standards addressing this issue. They can be used as guidelines for quantifying the relatively subjective variables that comprise this sub-dimension. The "Data Center Site Infrastructure Tier Standard: Operational Sustainability" from Uptime Institute [83] includes management and operations. The "Data Centre Operations Standard" from EPI addresses operations and maintenance requirements [84]. BICSI is currently developing the new Data Center Operations standard (BICSI-009) to be used as a reference for operations and maintenance after a data center is built.

4.5 Risk

Data center performance cannot be completely evaluated if the risks that may impact it are not considered. Optimization must involve risk, defined as potential threats that, if materialized, could negatively impact the performance of the data center.

The new family of data center metrics intends not only to measure performance for each process area, but also to associate it with its level of risk. That way, the user may implement actions to achieve the optimum performance and later adjust that performance to a tolerable level of risk, which may again deviate the metrics from their optimum performance. In the long term, if

variables remain unchanged, this model will lead to a stable equilibrium.

Risk of these sub-dimensions of performance, as well as the external risk which is independent of performance, are also measured through the use of metrics. They can be described as a causal system, where output depends on present and past inputs. An important strategy to reduce probability of failure is redundancy of resources, but this component may affect performance and costs [4].

4.5.1 Risk associated to performance

Table 5 presents general concepts for estimating the risk of the four identified sub-dimensions of performance.

Table 5. Risk related to performance

Risk	Concept
Productivity risk	Assessed as the downtime probability of occurrence times its impact. The probability is estimated using present and past data of downtime. Impact of downtime on productivity is calculated (e.g., the work that was not carried out with the required quality of service, as well as other indirect tangible and intangible costs associated with a failure). The impact would consider the cost of downtime [3].
Efficiency risk	Estimated with the ratio of processing utilization, IT equipment, physical space utilization, and IT equipment energy utilization, to their respective total capacities. Considers projected growth. When utilization is close to or at capacity, there is no room for growth, which means the risk that future projections will not be met is high. This directly influences performance.
Sustainability risk	Considers historic behavior of the different green energy sources, the percentage composition of each source, and their probability of failure.
Operations risk	Assessed by the operational risk, including documentation, planning, organization and human resources, maintenance, service level of agreement, and security. Analysis of historical data in order to estimate probability of failure due to improper operation in the areas identified, and its impact on performance.

4.5.2 External risk

Performance risks are not the only risks involved in the data center. There are other major risk factors that are external to the actual operation of the data center that must be considered in this analysis, namely site risk.

The authors of this paper have previously proposed a *data center site risk metric* [65] [85], which is a component of the comprehensive family of new data center metrics proposed in this paper. The methodology of the *site risk metric* helps identify potential threats and vulnerabilities that are divided into four main categories: 'utilities', 'natural hazards and environment', 'transportation and adjacent properties', and 'other'. The allocation of weights among each category is based on the significance of the impact of these factors on the data center operation [85].

The methodology quantifies the probability of occurrence of the events according to five pre-established levels of likelihood, and estimates potential consequences of each event using five pre-established levels of impact. It calculates the total risk level associated to the data center location by multiplying the probability of occurrence by the consequences or impact of each threat. That product is then multiplied by the respective weight.

Through this analysis, the different threats and vulnerabilities can be prioritized depending on the values of the probability of occurrence, impact, and the assigned weight. Understanding risk concentration by category can add value when analyzing mitigation strategies. This methodology provides a good sense of what the different risks and potential threats and vulnerabilities are for a data center site.

The *site risk metric* score is summarized into a time dependent function for a specific time instant [85]:

Data Center Site Risk Score =

$$\sum_{i=1}^{tc} \sum_{j=1}^{k_i} RL_{(i,j)} * W2_{(i,j)}$$

Where:

i: Threat categories (i = 1,…,t)

tc: Total number of threat categories (tc = 4)
(i=1: Services, i=2: Natural disasters, i=3: Transportation and adjacent properties, i=4: Other).

j: Specific threat (j=1,… k_i).

k_i: Number of threats for each category
(e.g., k_1=5, k_2=10, k_3=10, k_4=6).

$RL_{(i,j)}$: Risk level for the specific threat.

$W2_{(i,j)}$: Adjusted weight for the specific threat.

The risk level (RL): $RL_{(i,j)} = PO_{(i,j)} * I_{(i,j)}$

Where:

$PO_{(i,j)}$: Probability of occurrence of the specific threat.

$I_{(i,j)}$: Impact of the specific threat.

The adjusted weight (W2) for the threat: $W2_{(i,j)} = W_i * W1_{(i,j)}$

Where:

W_i: Weight of the specific category.

$W1_{(i,j)}$: Weight of the specific subcategory.

The data center site risk metric varies from 1 to 25, where 1 is the lowest level and 25 is the highest level of risk achievable for a specific site. The user must determine an acceptable level of risk, so that if the final score lies above that level, that particular location is not recommended.

4.6 New family of Data Center Metrics

The proposed new family of data center metrics can be summarized into a time dependent function, to be further developed. For a specific time instant, regardless of the correlation between different parameters, the data center score can

be defined as a function of different metrics, risks and weights:

Data Center Score =

$= f(P_1, P_2, P_3, P_4, R_1, R_2, R_3, R_4, R_E, W_1, W_2, W_3, W_4)$

$= f(P_i, R_i, R_E, W_i)$ with i=1,…,4

Where: P_i: Performance.

R_i: Risk of process.

R_E: External risk.

W_i: Weight of each category.

And sub-indexes: 1: Productivity.

2: Efficiency.

3: Sustainability.

4: Operations.

To the best possible extent, each value and weight assigned to each key indicator must be backed with enough support such as research or facts that lead to such conclusions. The outcome is a scorecard that assists in finding areas of improvement, which should be strategically addressed. The quality of information used before assigning each value is very important. Equally weighted data center scores are desired, since a data center must be productive, efficient, well managed and sustainable.

The new family of data center metrics scorecard can be taken as a decision-making trigger. It involves both technical and non-technical aspects, as failures and risks may not only be due to technical issues but also to non-technical ones such as human error. The metrics should measure parameters and processes. Given some premises, a data center may be ideal at a certain point in time, but when conditions change, that same data center may not be optimal. Depending on the final score calculated with the value of the dimensions, the scorecard would rank each data center on a scale to be defined, which allows for tangible comparisons between different data centers, or 'before and after' on the same data center. The multidimensional metric can also be transformed through different operators as a composite metric with just one value. Furthermore, the proposed metric has the possibility to incorporate new performance and risk measurements in the future, preventing it from becoming outdated.

4.7 Visualization tool

To confirm cross-comparability all the different indicators can be normalized. Each key indicator for performance can be presented in a scale interpreted in such a way that a higher value implies a more positive outcome, so minimum and maximum values correspond to the worst and best possible expected outcomes. Conversely, each risk value can be presented in a scale interpreted as a higher value implies a higher level of risk, and a more undesirable scenario. Figure 4 shows an example with indicators selected arbitrarily for illustrative purposes.

The graph (Figure 4) shows that the data center has a high productivity level, followed by efficiency, but its operations and sustainability indicators show greater room for improvement. Levels of risk associated to each category and site risk are low. Risk tolerance depends on the user, but working with this example, if we hypothetically set a maximum tolerable risk of 25%, actions would need to be implemented to reduce the risk related to productivity.

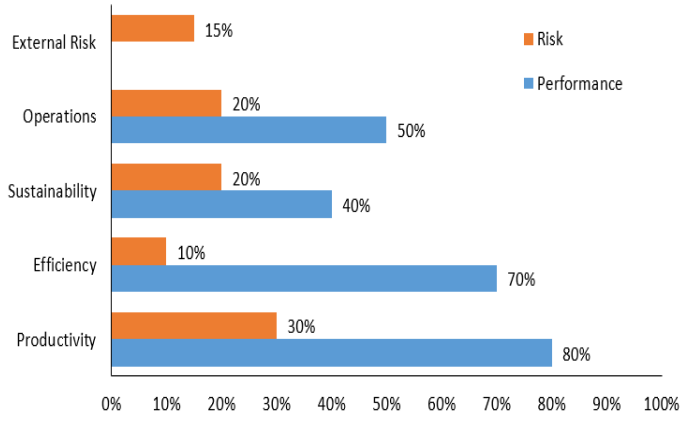


Figure 4: Data center multidimensional metric

Spider graphs can be generated to allow straightforward visual comparisons and trade-off analysis between different scenarios. This is very helpful when simulating or forecasting different strategies. It also enables clear reporting to stakeholders. Figure 5 shows an example with two different data centers, at different times that can involve specific strategies implemented.

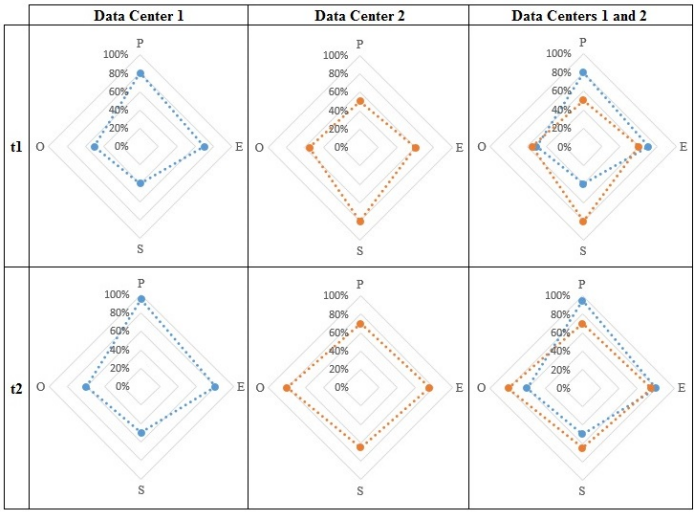


Figure 5: Data center multidimensional performance metric comparison (P: Productivity, E: Efficiency, O: Operations, S: Sustainability)

Edges of diamonds show measurement of the four dimensions of performance: productivity (P), efficiency (E), sustainability (S) and operations (O). The larger the diamond, the better the performance. It can be intuitively seen that in the scenario of time 1 (t1), data center 1 is more productive and efficient, but less sustainable and not so well operated, compared to data center 2. Over time (from t1 to t2), data center 1 has improved all its performance components, but data center 2 has worsened the

sustainability indicator and improved all other values. Risk can be analyzed in a similar spider graph, but for simplicity, it was not included in Figure 5.

4.8 Automation

Equipment needs permanent monitoring and maintenance to assure proper and efficient performance. Measurements should be automatic when possible as the automated metrics receive information directly from the different systems that quantify parameters. Real-time collection of relevant data is required for reliable metrics. Obtaining it is not a trivial task, especially in existing data centers that lack adequate instrumentation to collect the data [22], or if the related process cannot be easily automated. This underscores the need for new approaches to data center monitoring and management systems [86] [87]. Gathering the right data and understanding its nature is more important than simply collecting more data [88].

Real-time data can be gathered and updated automatically through a monitoring and management system [87] for parameters such as power consumption, temperature, humidity, air flow, differential air pressure, closure, motion, vibration, and IT equipment resource utilization. Improvements by some equipment manufacturers include the ability to directly access measurements of power, temperature, airflow, and resource utilization for each device. These measurements may include parameters such as the air inlet temperature, airflow, outlet temperature, power utilization, CPU utilization, memory utilization, and I/O utilization. Platform level telemetry can transform data center infrastructure management, allowing direct data access from the equipment processor [89]. Other parameters, such as some aspects of sustainability, operations and external risks, are more difficult to automate, since they require audits, human observation, evaluation, analysis, and the need to be periodically updated. All required data must be entered, automatically or manually, into a system in order to estimate all metrics and to visualize results clearly for decision-makers to undertake adequate actions.

5. Conclusions

The novel family of data center metrics as described in the paper provides a comprehensive view of the data center, using a multidimensional approach to combine performance (including productivity, efficiency, sustainability and operations) and risks (associated to performance and site risk). Using this approach, areas of improvement around which to create a strategy can be detected.

Given the mission critical nature of data centers, metrics must provide a holistic, yet quantitative, understanding of the data center behavior, in order to improve the utilization of all the resources involved. When issues identified by the metrics are addressed, processes can be optimized or moved closer to their desired point based on the vision for the data center.

Actions undertaken will impact the metric results in real time. When variables are re-measured, the result of the metrics should improve. As a result, new strategies may lead to modification of the overall metric as related to performance and risk values. This does not mean attempting to achieve the optimal performance. It is different from existing metrics in that it does not limit itself to measuring a value whose optimization can be automated, instead, this is only part of its scope. It aggregates measurements from multiple aspects to provide a global and objective vision to decide the extent to which optimization is desired. To recognize trends or predict future behavior, tools such as predictive analysis and machine learning are required. Machine learning considers algorithms that can learn from and make predictions on data.

The new family of data center metrics may contribute to standardize a process that eventually becomes a best practice. It may help to assess data centers, to compare them to each other, to make comparisons between different scenarios, and to provide a ranking of how the data center behaves. The outcome is a scorecard that will constitute a strong basis for decision-making.

6. Future Research

Further research needs to be conducted to assess and validate the new family of data center metrics proposed, based on a multidimensional approach and including performance and risk, in order to understand which parameters are most reasonable to use for each specific purpose.

Studying the correlation between the different metric scores and their associated parameters and risks, simulations of different scenarios can help to visualize how a change in parameters impacts risk, and likewise, how a change in risk factors affects metric results. Since there are currently no solutions available that include all proposed factors, new dynamic models and simulation tools are needed to validate and calibrate the metric. This would assist in implementing numerous data center strategies to improve metrics, for which a theoretical approach must be conducted. By tracking the proposed metric, the data center stakeholders will better understand data center performance and risk in a multidimensional view.

References

[1] M. Levy and D. Raviv, "A Novel Framework for Data Center Metrics using a Multidimensional Approach," *15th LACCEI International Multi-Conference for Engineering, Education, and Technology: Global Partnerships for Development and Engineering Education*. Boca Raton, FL, 2017.

[2] Ernest Orlando Lawrence Berkeley National Laboratory, "United States Data Center Energy Usage Report. LBNL-1005775," 2016.

[3] Ponemon Institute LLC, "Cost of Data Center Outages. Data Center Performance Benchmark Series," 2016.

[4] L. Minas and B. Ellison, *Energy Efficiency for Information Technology: How to Reduce Power Consumption in Servers and Data Centers*. Intel Press, 2009.

[5] "Data Center Design and Implementation Best Practices," ANSI/BICSI 002, 2014.

[6] S. Greenberg, E. Mills, and B. Tschudi, "Best Practices for Data Centers : Lessons Learned from Benchmarking 22 Data Centers," *2006 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, CA, pp. 76–87, 2006.

[7] D. Bouley, "Estimating a Data Center's Electrical Carbon Footprint. White paper #66," *APC by Schneider Electric*. 2012.

[8] U.S. Department of Energy. Office of Energy Efficiency & Renewable Energy, "10 Facts to Know About Data Centers," 2014. [Online]. Available: http://energy.gov/eere/articles/10-facts-know-about-data-centers.

[9] I. Ahmad and S. Ranka, "Chapter 43: Overview of Data Centers Energy Efficiency Evolution," in *Handbook of Energy-Aware and Green Computing*, vol. 2, CRC Press, 2012, pp. 983–1027.

[10] The Climate Group - Global eSustainability Initiative (GeSI), "SMART 2020: Enabling the low carbon economy in the information age," 2008.
[11] J. A. Laitner and K. Ehrhardt-Martinez, "Information and Communication Technologies: The Power of Productivity," American Council for an Energy-Efficient Economy, Washington, DC, 2008.
[12] ASHRAE Technical Committee 9.9, "2011 Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance," 2011.
[13] ASHRAE Technical Committee 9.9, "Data Center Networking Equipment – Issues and Best Practices," 2012.
[14] ASHRAE Technical Committee 9.9, "Data Center Storage Equipment – Thermal Guidelines , Issues , and Best Practices," pp. 1–85, 2015.
[15] "Energy Standard for Data Centers," ANSI/ASHRAE 90.4, 2016.
[16] "Green data centres – Part 1: Energy and environmental management systems," Singapore Standards SS 564, 2013.
[17] "Telecommunications Infrastructure Standard for Data Centers," ANSI/TIA-942-B, 2017.
[18] P. Turner, J. H. Seader, V. Renaud, and K. G. Brill, "Tier Classifications Define Site Infrastructure Performance," Uptime Institute, 2006.
[19] "Data Center Site Infrastructure Tier Standard: Topology," Uptime Institute, 2018.
[20] B. Howard, "H.R. 2126: The Energy Efficiency Improvement Act," 2014. [Online]. Available: http://www.usgbc.org/articles/hr-2126-energy-efficiency-improvement-act.
[21] T. Scott, "Data Center Optimization Initiative (DCOI). Memorandum," 2016.
[22] The Green Grid, "PUE™: A Comprehensive Examination of the Metric. White paper # 49," 2012.
[23] The Green Grid, "Green Grid Metrics: Describing Data Center Power Efficiency. Technical Committee White Paper," 2007.
[24] The Green Grid, "Green Grid Data Center Power Efficiency Metrics: PUE and DCIE. White paper # 6," 2008.
[25] U.S. Environmental Protection Agency, "Energy Star for Data Centers Scheduled Portfolio Manager Release on June 7 , 2010," 2010.
[26] W. Lintner, B. Tschudi, and O. VanGeet, "Best Practices Guide for Energy-Efficient Data Center Design," U.S. Department of Energy, 2011.
[27] N. Horner and I. Azevedo, "Power usage effectiveness in data centers: overloaded and underachieving," *The Electricity Journal*, vol. 29, no. 4. Elsevier Inc., pp. 61–69, 2016.
[28] V. Avelar, D. Azevedo, and A. French, *PUE™: A Comprehensive Examination of the Metric*. Atlanta, GA: ASHRAE Datacom Series, 2012.
[29] M. Andrea, "Data Center Standards. Data Center Size and Density. White paper DCISE-001." The Strategic Directions Group Pty Ltd, 2014.
[30] P. P. Ray, "The green grid saga - a green initiative to data centers: a review," *Indian J. Comput. Sci. Eng.*, vol. 1, no. 4, pp. 333–339, 2010.
[31] S. Sharma, C. H. Hsu, and W. C. Feng, "Making a Case for a Green500 List," *20th IEEE International Parallel & Distributed Processing Symposium (IPDPS) Workshop on High-Performance, Power-Aware Computing (HP-PAC)*. Rhodes Island, Greece, 2006.
[32] K.-D. Lange and K. Huppler, *Server Efficiency: Metrics for Computer Servers and Storage*. Atlanta, GA: ASHRAE Datacom Series, 2015.
[33] Sun Microsystems Inc, "Sun Sparc ® Enterprise T5120 and T5220 Server Architecture. White Paper," 2007.
[34] M. K. Herrlin, "Rack Cooling Effectiveness in Data Centers and Telecom Central Offices: The Rack Cooling Index (RCI)," *ASHRAE Trans.*, vol. 111, no. 2, pp. 725–731, 2005.
[35] M. K. Herrlin, "Airflow and Cooling Performance of Data Centers: Two Performance Metrics," *ASHRAE Trans.*, vol. 114, no. 2, pp. 182–187, 2008.
[36] C. D. Patel, R. K. Sharma, C. E. Bash, and M. Beitelmal, "Energy Flow in the Information Technology Stack : Coefficient of Performance of the Ensemble and its Impact on the Total Cost of Ownership." Hewlett-Packard Development Company, 2006.
[37] K. G. Brill and L. Strong, "Cooling Capacity Factor (CCF) Reveals Stranded Capacity and Data Center Cost Savings. White paper." Upsite Technologies, Inc., 2013.
[38] "2017 Standard for Performance Rating of Unitary Air-conditioning & Air-source Heat Pump Equipment," Air Conditioning Heating and Refrigeration Institute AHRI 210/240, 2017.
[39] "2016 Standard for Performance Rating of Computer and Data Processing Room Air Conditioners," Air Conditioning Heating and Refrigeration Institute AHRI 1360, 2016.
[40] "Method of Testing for Rating Computer and Data Processing Room Unitary Air Conditioners," ANSI/ASHRAE Standard 127-2012, 2012.
[41] The Green Grid et al., "Harmonzing Global Metrics for Data Centers Energy Efficiency," 2012.
[42] The Green Grid, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric. White paper # 32," 2010.
[43] The Green Grid, "Water Usage Effectiveness (WUE): A Green Grid Data Center Sustainability Metric. White paper # 35," 2011.
[44] The Green Grid, "ERE: A Metric for Measuring the Benefit of Reuse Energy From a Data Center. White paper # 29," 2010.
[45] The Green Grid, "Electronics disposal efficiency (EDE): an IT recycling metric for enterprises and data centers. White paper #53," 2012.
[46] J. R. Stanley, K. G. Brill, and J. Koomey, "Four Metrics Define Data Center Greenness," Uptime Institute, 2007.
[47] The Green Grid, "A framework for data center energy productivity. White paper # 13," 2008.
[48] The Green Grid, "The green grid data center compute efficiency metric: DCcE. White paper # 34," 2010.
[49] The Green Grid, "The Green Grid Data Center Storage Productivity Metrics (DCsP): Application of Storage System Productivity Operational Metrics. White paper #58," 2014.
[50] L. Newcombe, "Data centre energy efficiency metrics: Existing and proposed metrics to provide effective understanding and reporting of data centre energy. White Paper," BCS Data Centre Specialist Group, 2010.
[51] L. Newcombe, Z. Limbuwala, P. Latham, and V. Smith, "Data centre Fixed to Variable Energy Ratio metric DC - FVER: An alternative to useful work metrics which focuses operators on eliminating fixed energy consumption. White Paper," BCS Data Centre Specialist Group, 2012.
[52] I. Ahmad and S. Ranka, "Chapter 44: Evaluating Performance, Power, and Cooling in High-Performance Computing (HPC) Data Centers," in *Handbook of Energy-Aware and Green Computing*, vol. 2, CRC Press, 2012, pp. 1029–1049.
[53] eBay Inc, "Digital Service Efficiency - Solution Brief. White paper," San Jose, CA, 2013.
[54] D. King and S. Davies, "ACE Performance Assessment - Empowering the Data Center Operator. White paper," Future Facilities, 2014.
[55] J. Kaplan, W. Forrest, and N. Kindler, "Revolutionizing data center energy efficiency." McKinsey & Company, pp. 1–13, 2008.
[56] K. G. Brill, "Data Center Energy Efficiency and Productivity," Uptime Institute, 2007.
[57] Green IT Promotion Council, "Concept of New Metrics for Data Center Energy Efficiency - Introduction of Datacenter Performance per Energy [DPPE]," Japan, 2010.
[58] The Green Grid, "New Metric To Assess and Visualize Data Center Cooling," 2016. [Online]. Available: http://www.thegreengrid.org/~/media/Press Releases/The Green Grid_Press Release_Performance Indicator New metric_vFINAL.pdf.
[59] The Green Grid, "The Performance Indicator Assessing and Visualizing Data Center Cooling Performance. White Paper," 2016.
[60] Infrastructure Masons, "Engineering operations ratio. White paper." 2016.
[61] Infrastructure Masons, "Data Center Performance Index. White paper." 2017.
[62] R. Miller, "Class: New data center metric targets probability and risk," 2015. [Online]. Available: http://datacenterfrontier.com/class-new-data-center-metric-targets-probability-of-failure/.
[63] "IEEE Recommended Practice for Determining the Reliability of 7x24 Continuous Power Systems in Industrial and Commercial Facilities," IEEE Std 3006.7, 2013.
[64] Cushman & Wakefield, "Data Centre Risk Index," 2016.
[65] M. Levy and D. Raviv, "A framework for data center site risk metric," *IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*. New York City, NY, pp. 9–15, 2017.
[66] E. Jaureguialzo, "PUE: The Green Grid metric for evaluating the energy efficiency in DC (Data Center). Measurement method using the power demand," *2011 IEEE 33rd International Telecommunications Energy Conference (INTELEC)*. Amsterdam, 2011.
[67] F. De Angelis and U. Grasselli, "The next generation green data center a modified power usage effectiveness metric proposal for a CCHP natural gas or biogas fuelled architecture," *2015 IEEE 15th International Conference on Environment and Electrical Engineering, (EEEIC)*. Rome, pp. 907–911, 2015.
[68] R. Zhou, Y. Shi, and C. Zhu, "AxPUE: Application level metrics for power usage effectiveness in data centers," *2013 IEEE International Conference on Big Data*. Silicon Valley, CA, pp. 110–117, 2013.
[69] C. Fiandrino, D. Kliazovich, P. Bouvry, and A. Y. Zomaya, "Performance Metrics for Data Center Communication Systems," *2015 IEEE 8th*

International Conference on Cloud Computing. New York, NY, pp. 98–105, 2015.
[70] D. Schlitt and W. Nebel, "Load dependent data center energy efficiency metric based on component models," *2012 International Conference on Energy Aware Computing*. Guzelyurt, Cyprus, 2012.
[71] T. Wilde *et al.*, "DWPE, a new data center energy-efficiency metric bridging the gap between infrastructure and workload," *2014 International Conference on High Performance Computing & Simulation (HPCS)*. Bologna, pp. 893–901, 2014.
[72] M. Seymour and S. Ikemoto, "Design and management of data center effectiveness, risks and costs," *Annual IEEE Semiconductor Thermal Measurement and Management Symposium*. San Jose, CA, pp. 64–68, 2012.
[73] M. Wiboonrat, "Risk anatomy of data center power distribution systems," *2008 IEEE International Conference on Sustainable Energy Technologies, ICSET 2008*. Singapore, Singapore, pp. 674–679, 2008.
[74] J. Liu, Y. Zhang, Y. Zhou, D. Zhang, and H. Liu, "Aggressive resource provisioning for ensuring QoS in virtualized environments," *IEEE Transactions on Cloud Computing*, vol. 3, no. 2. pp. 119–131, 2015.
[75] S. Parida, S. C. Nayak, and C. Tripathy, "Truthful Resource Allocation Detection Mechanism for Cloud Computing," *The Third International Symposium on Women in Computing and Informatics (WCI '15)*. New York, NY, pp. 487–491, 2015.
[76] S. C. Nayak and C. Tripathy, "Deadline sensitive lease scheduling in cloud computing environment using AHP," *J. King Saud Univ. - Comput. Inf. Sci.*, 2016.
[77] M. Wiboonrat, "Life cycle cost analysis of data center project," *2014 Ninth International Conference on Ecological Vehicles and Renewable Energies (EVER)*. Monte-Carlo, Monaco, 2014.
[78] L. Tomas and J. Tordsson, "An autonomic approach to risk-aware data center overbooking," in *IEEE Transactions on Cloud Computing*, vol. 2, no. 3, 2014, pp. 292–305.
[79] X. Jin, F. Zhang, S. Hu, and Z. Liu, "Risk management for virtual machines consolidation in data centers," *2013 IEEE Global Communications Conference (GLOBECOM)*. Atlanta, GA, pp. 2872–2878, 2013.
[80] L. Yu, T. Jiang, Y. Cao, S. Yang, and Z. Wang, "Risk management in internet data center operations under smart grid environment," *2012 IEEE Third International Conference on Smart Grid Communications (SmartGridComm)*. Tainan, Taiwan, pp. 384–388, 2012.
[81] L. Yu, T. Jiang, S. Member, Y. Cao, and S. Member, "Risk-constrained operation for internet data centers in deregulated electricity markets," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 5, 2014, pp. 1306–1316.
[82] United Nations, "Report of the World Commission on Environment and Development: Our Common Future," 1987.
[83] "Data Center Site Infrastructure Tier Standard: Operational Sustainability," Uptime Institute, 2014.
[84] "DCOS - Data Centre Operations Standard," EPI, 2016.
[85] M. Levy, "Análisis de Riesgos y Selección de Localidad: Dos Pilares Imprescindibles [Risk analysis and site selection: two essential pillars]," *Datacenter Dynamics Focus*, pp. 32–34, 2014.
[86] M. Levy and J. O. Hallstrom, "A New Approach to Data Center Infrastructure Monitoring and Management (DCIMM)," *IEEE CCWC 2017. The 7th IEEE Annual Computing and Communication Workshop and Conference*. Las Vegas, NV, 2017.
[87] M. Levy and J. O. Hallstrom, "A Reliable, Non-Invasive Approach to Data Center Monitoring and Management," *Advances in Science, Technology and Engineering Systems Journal*, vol. 2, no. 3. pp. 1577–1584, 2017.
[88] J. Koomey and J. Stanley, "The Science of Measurement: Improving Data Center Performance with Continuous Monitoring and Measurement of Site Infrastructure," 2009.
[89] J. L. Vincent and J. Kuzma, "Using Platform Level Telemetry to Reduce Power Consumption in a Datacenter," *2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*. Folsom, CA, 2015.

Appendix A

The following table presents a summary of the reviewed existing data center metrics, classified by type with the main promoter. In some cases, the main promoter could not be identified. Must be noted that each metric uses its own definition for some terms (e.g., efficiency, productivity, performance, and risk), and for comparison purposes the same metric must be used.

Table A.1. Existing data center metrics

Metric	Type	Promoter
Power Usage Effectiveness (PUE) Data Center infrastructure Efficiency (DCiE)	Efficiency. Energy.	The Green Grid
Data Center Size	Efficiency. Space.	Data Center Institute Standards endorsed (AFCOM)
Rack Density	Efficiency. Rack. Space and energy.	Data Center Institute Standards endorsed (AFCOM)
Data Center Density (DCD)	Efficiency. Space and energy.	The Green Grid
Fixed and proportional overhead	Efficiency. Energy.	BCS Data Centre Specialist Group
HVAC effectiveness Airflow efficiency Cooling system efficiency Cooling System Sizing Factor Air Economizer Utilization Factor Water Economizer Utilization Factor Air temperature Relative humidity Rack Cooling Index (RCI) Return Temperature Index (RTI). Sensible Coefficient of Performance (SCOP)	Efficiency. Cooling system.	American Society of Heating, Refrigerating and Air-Conditioning (ASHRAE)
Coefficient of Performance of the Ensemble (COP)	Efficiency. Cooling system.	Hewlett-Packard
Cooling Capacity Factor (CCF)	Efficiency. Cooling system.	Upsite Technologies Inc.

Energy Efficiency Ratio (EER) Seasonal Energy Efficiency Ratio (EER)	Efficiency. Cooling system.	Air-Conditioning Heating and Refrigeration Institute
UPS load factor UPS system efficiency Lighting density.	Efficiency. Electrical equipment.	---
Compute Power Efficiency (CPE)	Efficiency. IT equipment (server).	The Green Grid
SPECpower_ssj2008	Efficiency. IT equipment (server).	Standard Performance Evaluation Corporation
SPECvirt_sc2013	Efficiency. IT equipment (server). Consolidation and virtualization.	Standard Performance Evaluation Corporation
SPEComp2012	Efficiency. IT equipment (server). Highly parallel complex computer calculations.	Standard Performance Evaluation Corporation
SPECweb2009	Efficiency. IT equipment (server). Web applications.	Standard Performance Evaluation Corporation
FLOPS per Watt	Efficiency. IT equipment (server)/	TheGreen500
TPC-Energy: TPC-C and TPC-E	Efficiency. IT equipment (server). Online transaction processing	Transaction Processing Performance Council
TPC-Energy: TPC-H and TPC-DS	Efficiency. IT equipment (server). Business intelligence or data warehouse applications	Transaction Processing Performance Council
TPC-Energy: TPC-VMS	Efficiency. IT equipment (server). Virtualized environment	Transaction Processing Performance Council
Vmmark	Efficiency. IT equipment (server). Virtualization platforms	VMWare
Storage performance	Efficiency. IT equipment (storage).	Storage Networking Industry Association, Emerald Program and the Storage Performance Council
Space Wattage and Performance (SWaP)	Efficiency. IT equipment (server).	Sun Microsystems
Idle-to-peak power ratio (IPR)	Efficiency. IT equipment (server). Idle consumption.	---
Linear deviation ratio (LDR)	Efficiency. IT equipment (server). Power linearity.	---
Green Energy Coefficient (GEC)	Sustainability. Alternative energy.	The Green Grid
Carbon Usage Effectiveness (CUE)	Sustainability. Carbon emissions.	The Green Grid
Water Usage Effectiveness (WUE)	Sustainability. Water usage.	The Green Grid
Energy Reuse Effectiveness (ERE) Energy Reuse Factor (ERF)	Sustainability. Energy reuse.	The Green Grid
Electronic Disposal Efficiency (EDE)	Sustainability. Decommissioned IT equipment.	The Green Grid
Site-Infrastructure Power Overhead Multiplier (SI-POM)	Sustainability. Site physical infrastructure overhead.	Uptime Institute
IT Hardware Power Overhead Multiplier (H-POM)	Sustainability. IT equipment efficiency.	Uptime Institute
Deployed Hardware Utilization Ratio (DH-UR)	Sustainability. IT equipment efficiency.	Uptime Institute
Deployed Hardware Utilization Efficiency (DH-UE)	Sustainability. IT equipment efficiency.	Uptime Institute
Free Cooling	Sustainability. Free cooling.	Uptime Institute
Energy Save	Sustainability. IT equipment hibernate.	Uptime Institute
Datacenter Performance per Energy (DPPE)	Performance. Sustainability. IT Equipment.	The Green IT Promotion Council (Japan)
Data Center Performance Efficiency (DCPE)	Performance. Productivity.	The Green Grid
Data Center energy Productivity (DCeP)	Performance. Productivity.	The Green Grid
Data Center compute Efficiency (DCcE)	Performance. Compute resources.	The Green Grid
Data Center Storage Productivity (DCsP)	Performance. Storage systems.	The Green Grid
Data Center Fixed to Variable Energy Ratio (DC-FVER)	Performance. Productivity. Wasted energy.	BCS Data Centre Specialist Group
Digital Service Efficiency (DSE)	Performance. Sustainability. Cost.	Ebay
Availability, Capacity and Efficiency (ACE)	Performance. Efficiency. Cooling system.	Future Facilities
Corporate Average Datacenter Efficiency (CADE)	Performance. Efficiency.	Uptime Institute
Data Center Energy Efficiency and Productivity (DC-EEP)	Performance. Efficiency. Productivity	Uptime Institute
Performance Indicator (PI)	Performance. Cooling system.	The Green Grid
Engineering Operational Ratio	Performance. Cooling system.	Infrastructure Masons

Data Center Performance Index	Performance. Availability, Efficiency. Environmental.	Infrastructure Masons
Class metric	Risk. Probability of failure.	MTechnology
Data Center Risk Index	Risk. Operations.	Cushman & Wakefield
Data Center Site Risk Metric	Risk. Site location. Operations.	M. Levy

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 241-253 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Pole-Changing Windings for Close Ratio and 1:N Ratio Using the 3//Y / 3//Y Method

Gabor Kovacs*

Ganz Electric Works, 1095 Budapest, Soroksari ut. 30-34. Hungary

ARTICLE INFO

Article history:
Received: 02 June, 2018
Accepted: 28 July, 2018
Online: 05 August, 2018

Keywords:
Induction motor
Pole changing 3Y / 3Y method
Pole changing 1:N

ABSTRACT

Requirement for pole-changing motors emerged at the very early times of asynchronous motor design. Different solutions have been elaborated and some of them are generally used. One of these solutions in use is the so called pole phase modulation method. The so called 3 Y / 3 Y pole-changing winding being subject of this paper belongs to this group. This paper presents a new approach to this method. A new, complete and comprehensive study is introduced, including features and, moreover, detailed design guidelines. The method is applicable for any pole combination for the so called close ratio without any further particular consideration. As the target of the paper is to provide solution for high power application, it does not deal with balancing methods, but as a novelty, calculates the effect of those disadvantageous phenomena (asymmetries, even harmonics) in order to enable the designer for evaluation and decision. The study also reveals as a novelty that the same method is suitable for wide ratio as well, moreover, a new solution for theoretically infinite ratio expressed in 1:N, with supply voltage system of N phase is proposed. Using this motor, a new and complete drive system has been proposed with subject pole changing motor as core element as the most appropriate arrangement of high power main marine propulsion drive.

1. Introduction

This paper is an extension of work originally presented at the International Conference on Electrical Machines 2018. organized by the World Academy of Science, Engineering and Technology [1].

The original work is extended in terms of

- much wider references are presented and deeply analyzed
- in case of close ratio,
    - the effect of the unavoidable even harmonics are not only generally analyzed but calculation methods are given as well; the limits for omitting balancing measures are clearly given and are supported by calculations
    - calculation methods are given to support the work of a designer of such motor (such as: winding pitch, maximum fundamental torques, air gap induction, harmonic torques, current imbalance).
- in case of wide ratio,
    - the necessity of NxN-times phase supply on the lower pole number is clearly explained by voltage vector diagram; winding pattern also for 1:5 is given demonstrating how to create the winding of any ratio for 1:N by applying N-phase supply system and N//Y / N//Y connection
    - supply strategies are given supported by calculation.

2. Pole changing motors

The objective of motor designers for industrial use was, from the beginning, to provide two speeds in a single motor instead of one speed only.

There are two main methods for pole changing of motors which are most generally used. The first method creates two independent windings in the stator, which is basically two motors in one: both windings can be ideally designed for their respective speeds in the meaning that the number of turns per slot may be chosen freely corresponding to the number of the pole. Due to the fact, however, that only one of the windings works at a time, the size of such motors is much bigger than the size of a corresponding single speed

*Gabor Kovacs, Ganz Electric Works, Email gabor.kovacs.ganz@gmail.com

motor. The method is generally used for close ratio, practically any pole combination can be implemented [2].

The second method is based on changing the connection of parts of a winding called pole amplitude modulation (PAM motor): a part of the winding is connected in a to-and-back direction in order to achieve the basic and then the modulated pole number. Although the entire winding is working on both speeds at any given moment, two inherent disadvantages cannot be avoided: a number of additional terminals are required, therefore the drive must be fitted with a specially designed switching apparatus, and the magnetic field/wave in the airgap is always considerably distorted on the modulated speed. With this method, the possible pole combinations are limited. The method is used both for close ratio as well as for wide ratio.

Another group of methods is the one so called pole phase modulation. With this method, subdivided parts of the winding belong to one phase for one pole number while some of them will be moved to another phase for forming the other pole number. Intensive research has been performed on this field so far. Although the method is suitable for both close and wide ratio, the research is more focusing on the wide ratio nowadays.

Serious proposals for wide ratio might have started with the pioneer works of [3] and [4] both research groups working parallel to each other, reaching almost the same result, proposing 1:3 ratio pole changing winding with ideal (inherently balanced) characteristics. They admit, however, "it is not suggested that this phenomenon is likely to have any application but is of interest as an instructional demonstration". [5] gives design criteria for pole changing motors as a summarizing work of the results till that time. 3 parallel connection was proposed at an early time by that milestone work of [6], with advantageous 6 terminals only. Its strategy is, however, completely different from that given in present paper. Yet, it was expressly stated by them that the method is suitable for close ratio only. Further, each pole combination required special considerations to implement, sometimes with coil omission; and for low power only. Current imbalance between parallel branches was just mentioned, but not discussed, not explained, even not calculated. [7] states to have developed a computer program for any pole combination of close as well as wide ratio prevailing on the previous results of the technique. 1:4, even 1:7 ratio machines were tested but at laboratory size only. [8] gives again a special solution for 1:4 ratio, again for low power. In [9], a 1:3 pole ratio motor was developed for a 9-phase supply on the lower pole number. Moreover, they studied and solved the smooth transition from one pole to the other by the control of the inverter supply, marking out a direction for future research for others. [10] and [11] were dealing with the same motor. The connection for a pole number to the other and for a phase number to the other are not made within the motor but by the control of the 9-phase supply. Therefore, the solution is rather a pole changing system than a pole changing winding. [12] is again a paper with summarizing value by giving general winding design rule for the pole phase modulation. Control algorithms were also developed. [13] and others not listed here are strongly working on eliminating imbalance of the winding for close ratio, proposing successful balancing methods.

These methods, however, although being suitable for low power but cannot be applied for high power motors (e.g. different turn numbers per slot). [14] is one example for the seemingly recent direction of research where as many as 5 different pole numbers are integrated in one motor. The solution is again such that the connection (and the pole changing at all) is not made within the motor but by the control of the supply (each slot being controlled by its own inverter).

Generally, it can be stated that the works are related to low power, laboratory size and they propose particular solution only. Some solutions are proposed for close ratio only while the rest only for wide ratio, respectively. Also the help of the supply is required in many cases for implementing the pole changing effect.

Approach of each paper (and of those not listed here) is different from that presented in this paper; work presented in [15] and [16] was not taken by anybody into consideration.

3. 3//Y / 3//Y Method

Belonging to the family pole-phase-modulation, another alternative is the so-called 3//Y/3//Y method. It is a real pole changing winding. Advantages of this solution are that the entire winding is working at any given moment and only 6 terminals are needed. At the same time, however, certain characteristics may be unfavorable. There may be certain asymmetries and additional even harmonics values. Therefore, calculation methods shall be given for those undesirable phenomena, in order to evaluate their importance.

As the 3// connection is inherently suitable for high power and balancing measures proposed by others are hardly feasible on that field, this paper will not deal with them. Instead, calculation methods will be introduced in order to enable the designer to evaluate those unfavorable phenomena.

It will be shown that both close ratio as well as wide ratio pole changing winding can be implemented by this single method.

4. Basic Idea behind the 3//Y/3//Y Method

Figure 1 (see [15] and [16]) shows a pole-changing connection for 4/6 pole. Instead of the usual 60° phase belt, such connections are prepared with a 120° phase belt and two-layer winding. Figure 1 (a) and (b) shows the upper layer for 4-pole and 6-pole connections, respectively. Figure 1 (c) shows the necessary subdivision into parts being smaller than q slot (slot number per phase per pole) on 4 pole.

The figure indicates the relationship between those subdivided parts of the winding and phases, indicating a 4-pole connection in one case (d) and a 6-pole connection in the other (e). It is clear that certain winding parts belong to a certain phase for 4-pole and to another phase for 6-pole.

In subject case, winding group 1, 4, 7 remains in the same phase; however, all the rest is moved to another phase. Because the groups are connected into series, 15 terminals and an appropriate switch are required.

It is clear that the elements of groups 2, 5, 8 as well as 3, 6, 9 shall be identical among each other; otherwise they could not be interchanged with each other.

Considering those three groups as a whole, it becomes clear that group 1, 4, 7, group 2, 5, 8 and group 3, 6, 9 are also identical with each other allowing them to be connected in parallel.

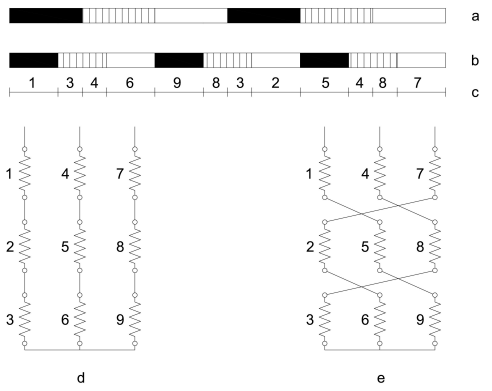


Figure 1. Classical pole changing winding according to [15] and [16]
(a) upper layer for 4-pole connection
(b) upper layer for 6-pole connection
(c) subdivision of the winding
(d) connection of winding for 4-pole
(e) connection of winding for 6-pole

This consideration opens a new way for a much simpler solution.

See connection pattern on Figure 2.

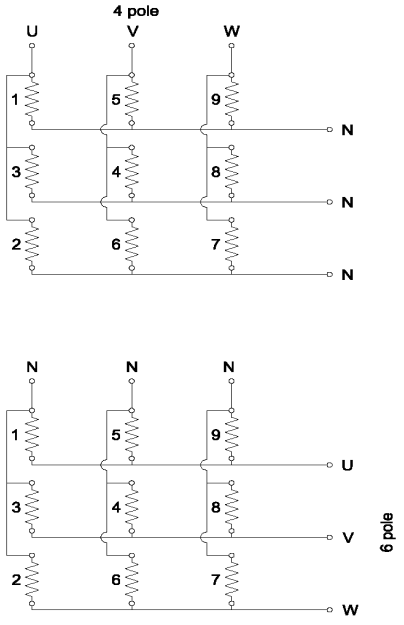


Figure 2. Pole changing winding by 3//Y / 3//Y method

5. Application of the method for close ratio

5.1. Phase Imbalance between Groups to be connected in Parallel

Such question does not arise for the original series connection, as just because of the series connection, the system is inherently balanced. The same is not evident, however, for the parallel connection. Therefore, Figure 1. and Figure 2. shall be considered again showing the origin and the derivation of the method.

The method of derivation for the 6-pole connection will be analyzed first. It is clear that always full poles are connected into the final scheme (no further subdivision for these) and these full poles are obviously always in the right phase angles. Therefore the 6-pole connection is always and inherently balanced.

In another case when 6 pole is the lower pole number, its full poles are, however, necessarily subdivided. Yet these subdivided parts (taken from different poles) are in the end always connected into series in a way as to always create full poles again see and consider Figure 5a. Generally, it stands that if the pole number is an integer multiplier of 6, it is always balanced, regardless of whether it is the higher or the lower pole number.

Returning to the derived 4-pole connection and in order to make the explanation more transparent, some of the already subdivided parts "are further cut" in order to better illustrate the phase angle of the voltage induced in them see Figure 3a. Then, in Figure 3b. the resulting induced voltage pattern are provided for each 3-phase group to be connected into parallel.

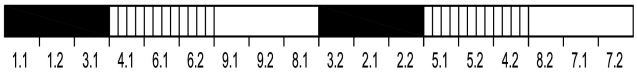


Figure 3a. Sketch for 4-pole connection serving as reference for Figure 1a.

The voltage values of the 3 groups to be connected in parallel are not balanced. It is clear that the vector difference between the supply and the induced voltage of each group (being always 20°) will drive equalizing currents in each group. In order to extend the picture the same pattern is provided for 8 pole as well. Figure 4. indicates the final result only.

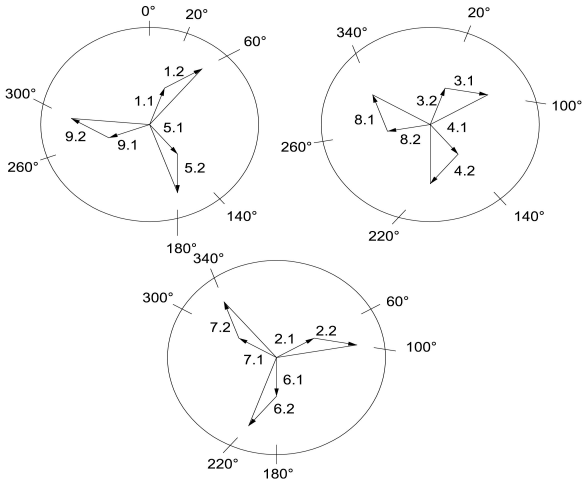


Figure 3b. Voltage vector pattern for 4-pole connection

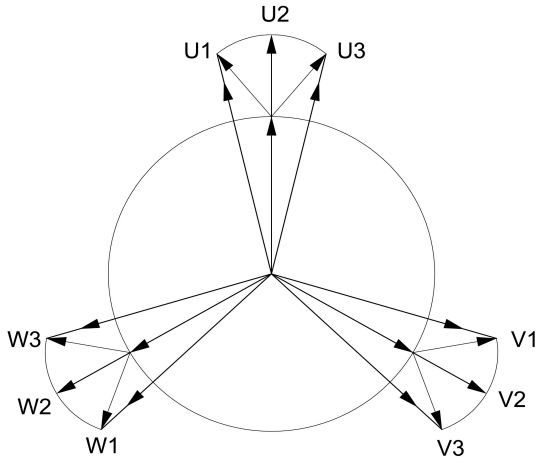


Figure 4. Voltage vector pattern for 8 pole connection acc. to Winding Scheme on Figure 5b.

As seen the imbalance is less. The reason for this is that, in case of 4 poles, all poles are subdivided into 3 parts that results necessarily in no right phase angle. At 8 poles, however, regardless of whether it is the higher or the lower pole number, the "first 6 poles" are always being put together creating full poles (and symbolized by those 3 vectors in the circle being 120° from each other), and only the "last 2 poles" are subdivided in an undesirable way (symbolized by those 3 vectors starting from those previous ones, always 40° from each other). Therefore, the 8 pole connection contains definitely fewer imbalances. The same consideration applies for higher pole numbers. Both Figure 3b and Figure 4 show the vectors in a true scale.

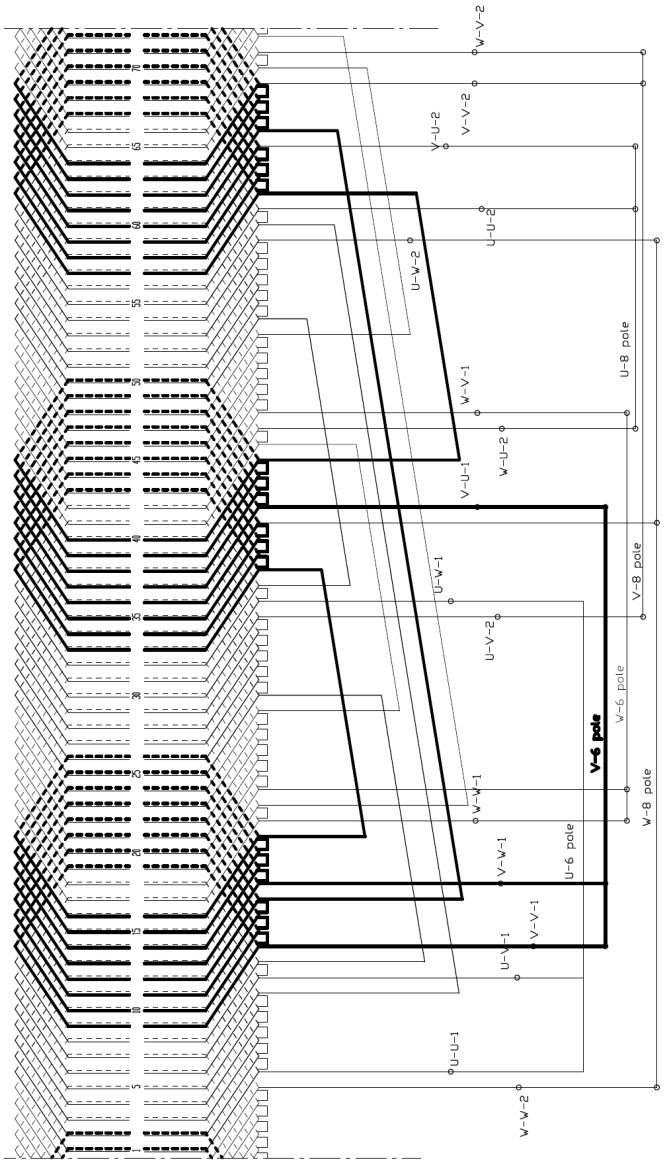


Figure 5a. Winding Scheme when supplied on the terminals marked by 6 pole

The induced voltages as well as (obviously) the supply voltage of each group constitute a symmetrical 3-phase system. It is the vector difference between them which drives the equalizing currents. Therefore no negative sequence currents will arise. The equalizing current will be appr. perpendicular to the magnetizing current in group 1 and group 3, while it will be parallel and in the sense against the magnetizing current in group 2. Then, the superposition method shall be used for the resulting current (and resulting air-gap induction) in each group. Local irregularities (copper losses, air gap & iron core induction) arise in this way. A short approximate calculation method will be given for the equalizing currents in the Appendix.

5.2. Actual implementation

An actual implementation by showing both layers of the winding is presented on Figure 5. as an example and demonstration. It shows how the winding of a real motor looks like in the high power industry, not in a laboratory. From now on a 6/8-pole winding will be taken as an example for further considerations.

Figure 5a shows the result supplied at the terminals marked for 6-pole, while Figure 5b shows the result supplied at the terminals marked for 8-pole.

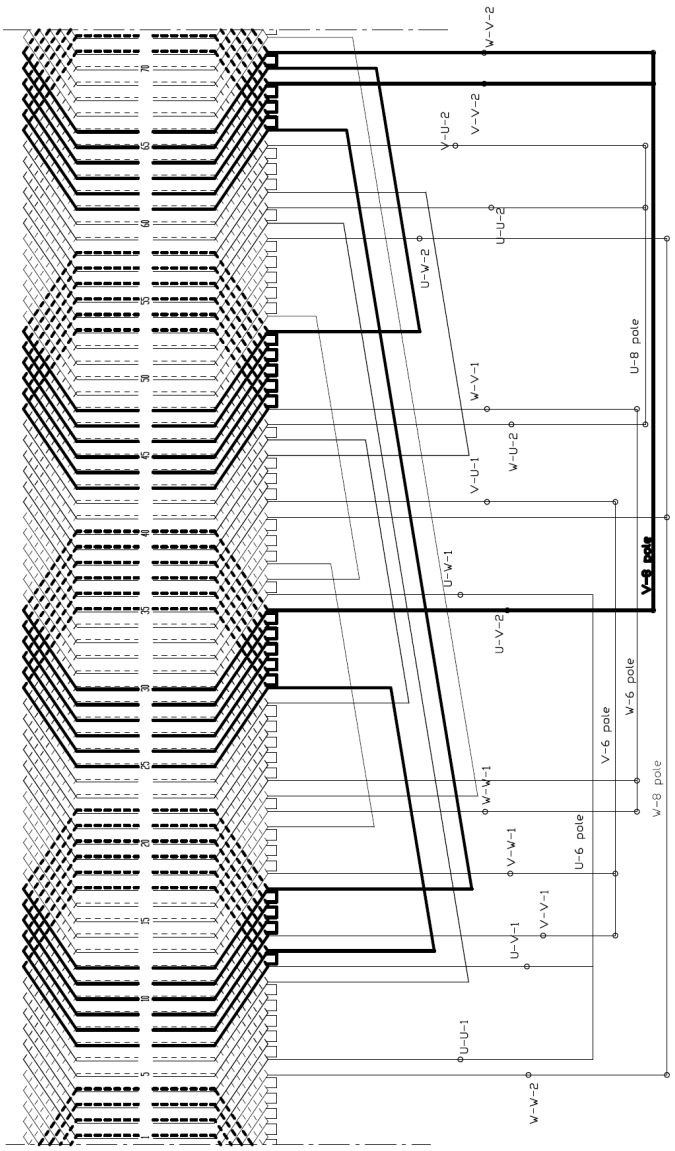


Figure 5b. Winding Scheme when supplied on the terminals marked by 8 pole

There are three terminals for both pole numbers: one group is used for power supply for 6-pole, whereas the other three terminals are used as the star points, therefore these latter remain idle. The opposite is true for 8-pole: the other group of three terminals, used

earlier as star points, is used now for power supply, whereas the former group remains idle as star points. Those terminals, serving actually for star-points, do not need to be short-circuited.

The basic connection scheme in [15] and [16] see Figure 1. implies the concept of subdivision of the winding (=of the poles) and then the re-grouping as well. It is presented there as a particular solution, only for 4/6 poles. By understanding the concept, however, a winding for any pole combination can be created with no limitation. Endwinding connections are "longer" than those of usual single speed motors but note that each of them shall be dimensioned on the 1/3 of rated current only.

5.3. Design Aspects

5.3.1. Stator slot number

The possible stator slot numbers are considered for this method now. Because the slots are subdivided into nine groups, the slot numbers must be integer multiple of 9, no matter which are the pole combinations.

The best is to apply the rule for slot numbers in case of two independent windings [2]. Three-phase motor slot numbers are always integer multiple of 3. Slot numbers possible generally for pole changing motors with two independent windings are integer multiple of a further coefficient of 3 if any of the pole numbers is integer multiple of 6, so the slot numbers suitable for pole changing winding will be inherently integer multiple of 9 in such cases. For the rest of combinations like 8 – 10, 10 – 14, 14 – 16, etc., that above resulting slot number must be multiplied by 3.

5.3.2. Rotor slot number

The subject method does not imply any new requirement regarding rotor slot number, therefore traditional considerations may be applied. Yet, because of the even harmonics, in order to reduce the magnitude of their counteraction with other fields, rotor slot numbers higher than that of the stator may be recommended.

5.3.3. Winding pitch

The main (or perhaps the only) tool for the designer having in hand: the winding pitch. This fact has gained little consideration so far, it is for this 120° phase belt, however, of basic importance. Again, see Figure 5a for the lower pole number and then Figure 5b for the higher pole number.

In case if the pitch would be 1 – 13, this would mean full pitch for 6 pole, the arrangement is symmetrical, that is no even harmonic, the consecutive poles (N – S ...) are equal. It is clear at the first glimpse that the picture for the higher pole number cannot be symmetrical at the same time. In this case, pitch is 240° for the 8-pole, therefore (considerable) even harmonics will evolve.

If the slot pitch is, however, not 1 – 13 but 1 – 10, the situation is just the opposite: the winding of 8 pole is symmetrical (= no even harmonic) and 6-pole is (extremely) not symmetrical. If the pitch is in between, like 1 – 11 as on Figure 5. indeed, both are non-symmetrical but just on a moderate extent.

It is not usual to choose a pitch over 180° for the usual 60° phase belt winding; but this value is exceeded at such pole-changing winding for the higher pole number. The right choice of pitch is the key in designing such pole-changing motor where the actual drive and operation must always be taken into consideration. For example, if the higher pole number, that is lower rated speed, is applied only for starting and the standard operation always occurs on the lower pole number, then selecting a pitch which is symmetrical on the lower pole number (= higher operating speed) is recommended.

The following considerations will help in the decision:

- Max. 2nd harmonic will evolve: pitch degree 270°
- Disappearing 2nd harmonic: pitch degree 180°
- Max. 4th harmonic evolves: pitch degree 135° or 225°
- Disappearing 4th harmonic: pitch degree 180° or 270°

Based on this, the target is to choose a pitch that gives a degree between 135° and 225° on both pole numbers. This is possible for pole numbers being "close" to each other, e.g. 6 – 8, 8 – 10, 10 – 12, etc.. On those pole numbers which are "far" from each other as 4 – 6, 6 – 10, etc., only those pitches are possible where the pitch degree for the lower pole number is between 135° and 225° and the pitch degree for the higher pole number is over 225°. Vicinity of 225°, however, must be avoided in any case.

In order to give proposals in this respect the designer shall make some calculations first; it can be analyzed only after whether the right or even ideal pitch might be determined if such exists at all.

5.3.4. Harmonic Phenomena

In this chapter the winding factors will be determined first and then conclusions will be drawn regarding harmonic leakage, attenuation and harmonic torques. For reasons of generalization, an infinite number of slots will be applied. Then, as soon as the actual slot number is known, the known formulas may be used.

5.3.4.1. Distribution factor

ν =	Distribution factor
1	$\sqrt{3}\cdot 3/2\pi$
2	$-\sqrt{3}\cdot 3/4\pi$
3	
4	$\sqrt{3}\cdot 3/8\pi$
5	$-\sqrt{3}\cdot 3/10\pi$
6	
7	$-\sqrt{3}\cdot 3/14\pi$
	…….

(1)

5.3.4.2. Chording factor:

$$\sin(\nu\cdot\pi/2\cdot y/\tau) \quad \text{where } \tau \text{ is full pitch} \tag{2}$$

Based on the above the differential leakage and the attenuation factor can be calculated acc. to the theory but with even harmonics also involved.

5.3.4.3. Differential Leakage Coefficient:

$$\sigma_1 = \sum_{\nu}\left(\frac{\xi_\nu}{\nu\cdot\xi_1}\right)^2 \tag{3}$$

For $\nu = 2, 4, 5, 7, 8, 10, 11, 13 \ldots$

5.3.4.4. Attenuation Factor:

$$\Delta = 1 - \frac{\sum_{\nu} \eta_\nu^2 \cdot \left(\frac{\xi_\nu}{\nu \cdot \xi_1} \right)^2}{\sum_{\nu} \left(\frac{\xi_\nu}{\nu \cdot \xi_1} \right)^2} \tag{4}$$

where

$$\eta_\nu = \frac{\sin\left(\nu \frac{p \cdot \pi}{N_2} \right)}{\frac{p \cdot \pi}{N_2}}$$

with the N_2 rotor slot number that cannot be infinite here.

5.3.5. Harmonic Torque:

It is not within the scope of this paper to take a position between the different approaches on how to calculate harmonic torques. For the sake of good transparency, the method given in [17] p. 112. (240a) will be followed here:

$$M_\nu = \frac{3}{2} \frac{U}{I_{m0}} \frac{\nu}{\omega} \frac{I_k^2}{1 + \tau_d} \frac{\xi_\nu^2}{\nu^2 \cdot \xi_1^2} \tag{5}$$

Where

U – phase supply voltage
I_{m0} - magnetizing current (fundamental)
I_k - short circuit current (fundamental)
τ_d - differential leakage coefficient

For obtaining a more simple formula that gives quick values for comparison, some further formulas will be introduced and certain approximation will be made as follows:

$\tau_d \to 0$

$M_{\max} = \frac{3}{2} \frac{U^2}{X_s} \frac{1}{\omega}$ fundamental

$I_k = U / X_s$ - with X_s stray reactance

$I_{m0} = U / X_m$ - with X_m magnetizing reactance

By inserting the above three approximate relations into (5) and by rewriting the following formula will be obtained:

$$M_\nu = M_{\max} \frac{X_m}{X_s} \frac{\xi_\nu^2}{\nu \cdot \xi_1^2} \tag{6}$$

If both X_m and X_s will be put in p.u. a quick estimation will be obtained for the maximum torque of the νth harmonic in relation to the maximum torque of the fundamental wave.

This formula must be applied for $2p_1$ and $2p_2$ pole number separately because also p.u. values of X_m and X_s are different on each pole number.

From parasitic torque point of view the 4th harmonic is the critical one, because its rotation is in the motor range, in other words the sense of rotation is the same as that of the fundamental field. The 2nd harmonic is rotating opposite sense therefore it has no influence on the motor range unless its break-down torque is extremely high and its effect (although acting in brake-range) "reaches" the motor range. Anyway, harmonics' torque may be reduced linearly by reducing magnetizing reactance (= increasing no-load current) see [17] p. 112, equ (240a): which may be a reasonable compromise. 4th harmonic torque has less effect at a drive with parabolic counter-torque like pump, fan etc. Corresponding effect on motor derating k3 is discussed under item *5.3.9.* Derating. of this chapter.

From differential leakage point of view, however, the 2nd harmonic is the critical one, because its influence on the differential leakage may be dramatic, although having highest attenuation. Therefore, for example, pitch degree of 270° is not recommended, although 4th harmonic would disappear but 2nd harmonic is just on maximum.

In case of critical situation regarding too high 4th harmonic, the following method can be applied: the "last" slot or slots in each phase belt will be left without winding. By doing so the phase belt will be less than 120°, or with other words, it gets one-step closer to phase belt 90°. If 90° phase belt were reached, there would be no 4th harmonic – as known from the theory.

Obviously, each "empty" slot requires further derating of the motor, therefore just a moderate use of this method is recommended, also because further harmonics may be generated at the other pole number; see k4 under item *5.3.9.* Derating. of this chapter. Calculations show that practically no change of differential leakage is expected by doing so.

5.3.6. Maximum Fundamental Torque on p_1 and p_2 Pole Numbers, Air-gap Induction

The fundamental air gap flux will be determined first and then the maximum torque will be estimated on both pole numbers. The expression of voltage and flux will be written as known:

$$U = \sqrt{2} \cdot \pi \cdot f \cdot N \cdot \xi_1 \cdot \Phi \tag{7}$$

with f – supply frequency
Φ – flux per pole
N – number of turns per phase.

As the N number of turns per phase is definitely identical on both pole numbers, it follows that $\xi_1 * \Phi$ is necessarily identical also. Because the maximum torque is proportional to

$$M_{\max} \approx p * \xi_1^2 * \Phi^2 \tag{8}$$

the following will be obtained:

$$M_{\max p2} \approx p_2/p_1 \cdot M_{\max p1} \tag{9}$$

This means the absolute value of maximum torque on the higher pole number is definitely higher than that on the lower pole number. The motor operation may be called as "constant power operation".

Furthermore, with

$$\Phi = D \cdot \pi/2p \cdot L \cdot B$$

with D – air gap diameter
L – equivalent length of iron core
B – air gap induction mean value

Inserting it into (7) the following formula is obtained:

$$B_{p2} = (p_2/p_1) \cdot (\xi_{1p1}/\xi_{1p2}) \cdot B_{p1} \tag{10}$$

As important information for the designer, it is established that the air-gap induction on the higher pole number is (much) higher than that on the lower pole number.

5.3.7. Ideal pitch

Because the pitch is practically the only tool in the hand of the designer the question arises whether the optimum or "ideal" pitch may be found or not. Figure 6. will help in the consideration.

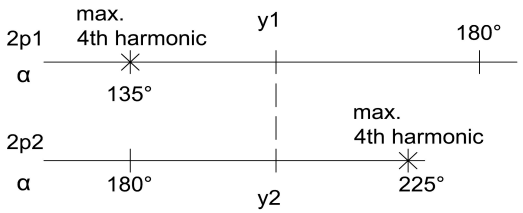


Figure 6. Sketch for explaining "ideal" pitch

The sketch indicates the electrical angles of both pole numbers, around 180° as "full pitch", as well as those dangerous pitches which should be avoided. The coincidence of y_1 and y_2 demonstrates that the mechanical pitch is necessarily identical. Therefore

$$y_2 = p_2/p_1 \cdot y_1$$

While moving y_1 towards 180°, the harmonic content (incl. 4th harmonic) on the lower pole number will always be more and more advantageous, but, at the same time it will always be more and more disadvantageous on the higher pole number until it reaches unacceptable extent, and vice-versa see also in 5.*3.3.* above.

The right pitch shall be somewhere in between. A possibility to choose y_1 and y_2 in a way that

$$180^0 - y_1 = y_2 - 180^0$$

At such choice the winding factors of all harmonics as well as the fundamental will be identical on both pole numbers.

The resulting pitches are

$$y_1 = \frac{p_1}{p_1 + p_2} \cdot 360^0 \quad y_2 = \frac{p_2}{p_1 + p_2} \cdot 360^0 \tag{11}$$

Harmonic torques p.u. value as given in (6), however, will still not be identical because they are further dependent on the ratio of X_m and X_s, and those are different for each pole number. Further consideration is necessary in this respect.

Passing over detailed deduction it is obtained that

X_m is proportional to $(1/p)^2 \cdot \xi_1^2$

while X_s however, remains approximately constant (more precisely it reduces slightly).

It is obtained at the end that the ratio X_m / X_s valid for the lower pole number will change when connected to the higher pole number according to the following way:

$$(X_m/X_s)\big|_{p2} \approx \left(\frac{p_1}{p_2}\right)^2 \cdot \frac{\xi_{1p2}^2}{\xi_{1p1}^2} \cdot (X_m/X_s)\big|_{p1} \tag{12}$$

This expression shall be inserted into (6) for direct calculation of harmonic torques on the higher pole number. It is underlined that the formulas give the harmonic torque referred always to its own maximum fundamental torque (in p.u.).

Targeting equal 4th harmonic torque (p.u.) on both pole numbers as a good compromise, the above formula tells that y_1 and y_2 according to (11) will not provide the expected results as the harmonic torques will be much less for the higher pole number. Therefore, y_1 may be moved a bit towards 180° by reducing chording on the lower pole number. Because of the non-linearity (sin() in both winding factors, ξ^2 in the torque expression) this "ideal" pitch cannot be given in a definite formula.

y_2 – obviously - must not be moved too close to 225°.

For a full picture, the calculations by the formulas in *5.3.4.3* and *5.3.4.4* give that

- this "targeted" pitch (although for infinite slot numbers) will result in such "high" magnitude of differential leakage coefficients as given in the literature for q = 3 or so. This is due to the presence of (low-order) even harmonics.
- at the same time it results in lower attenuation factors than given in the literature due to the strong attenuation effect of the rotor just on the low-order even harmonics.

Expected approximate harmonic torques for 6/8 pole at that chording for best compromise:

6 pole		8 pole	
M_4 / M_{max}	0.16	M_4 / M_{max}	0.16
M_7 / M_{max}	0.007	M_7 / M_{max}	0.006
M_{10} / M_{max}	0.023	M_{10} / M_{max}	0.001

This example represents the limit of application in respect of pole ratio (min. 0.75) without balancing measures.

5.3.8. Parallel Connection of the Winding

For high power motors, further parallel connection of the winding may become necessary. It is, unfortunately, not possible for the smallest pole numbers, e.g. for 6-8 pole or such. It is possible, however, for higher pole numbers: for 12-16 pole further 2// connections, for 18-24 pole further 3// connections become possible, for example.

5.3.9. Necessary Derating of the Basic One-Speed Motor

Based on the above considerations, the necessary derating of such double-speed motor can be defined compared to the basic single-speed motor. To put it differently: considering a certain size of a single-speed motor with P [MW] power and $2p_1$ pole number, what can be the power P1 of such double-speed motor with the same lower pole number $2p_1$. P2 belonging to $2p_2$ (the higher one) is not important, because the load at the higher pole number is always considerably lower.

There are 4 factors for necessary derating:

$$P1 = k1 * k2 * k3 * k4 * P \tag{13}$$

Each coefficient is < 1, according to the followings:

k1 derating due to the ratio of the winding factor for 60° and 120° belt and for the pitch. The belt factor ratio is always the same: 0.827 / 0.955 = 0.866 which is the value of inherent derating necessity derived directly from the theory of this method. Then, this should be multiplied by the ratio of the actual (lower pole number) pitch shortening factor compared to the usual 5/6 shortening factor. This later ratio is usually again less than 1.

k2 derating due to excess loss caused by the balancing currents due to imbalance of stator phase angles. This only applies in case the lower pole number is not integer multiple of 6, otherwise k2 = 1

k3 derating necessity due to eventual excess loss caused by eventual necessity of increasing no-load current, which causes the rated current to increase because of the decreased cos φ. For further details see Harmonic Torque *5.3.5.*

k4 derating necessity in case where phase belt should be less than 120°, again see Harmonic Torque *5.3.5.*

Overall derating necessity is expected to be 75 – 80% (example: a 6 MW single speed motor size may be loaded by 4.5 – 4.8 MW if wound for 3// Y / 3 //Y for any pole number).

5.4. Conclusion, proposal for application

Any pole number combination can be implemented by the prescribed method.

Based on the voltage imbalance and the calculation of the harmonic torques it can be established, however, that the method may be applied

- for 6 poles and above and
- for pole ratios of 0.75 and higher (very close ratios)

without any balancing measures in the winding.

The method is recommended especially for driving such machines as pumps, fans etc.

It is underlined that such pole combinations like 4-6, 6-10 etc. can be implemented also. The high unfavorable effects of the even harmonics and phase imbalance appearing here may be eliminated by balancing measures.

6. Application of the method for wide ratio

6.1. General Considerations

Now it will be investigated as a new field of research whether the method can or cannot be applied for pole numbers being very far from each other, generally called wide ratio. The easiest investigation involves a 1 : N ratio where N is integer.

If N is even, such as 2, 4, 8, etc. and symmetrical pitch (=full pitch) is chosen on the lower pole number, as logical first approach for avoiding even harmonics, there will be no excitation on the higher pole number (it is easy to understand that U and –U are in the same slot and therefore magnetomotive force of each slot is zero). As a consequence, such motor simply does not work. If, however, y_1 will be moved to either direction, but preferably towards lengthening, the machine will work on both pole numbers, but even harmonics appear also on both pole numbers; the calculation method can be obtained again from the previous chapter.

If N cannot be divided by 3 such as 4, 5, 7, 8 etc. the 3// connection cannot be implemented at all, because the 3 groups cannot be identical, as one phase is always and considerably different from both the other phases.

If N = 3 (3, 9 etc.), however, like a 6 – 18 pole changing winding, the result is absolute ideal, no even harmonic will appear on either pole number, therefore the motor's behavior on both pole numbers is identical to the corresponding single-speed motor.

This statement may be supported very easily.

If "full pitch" will be chosen on the lower pole number, as the logical first approach for avoiding even harmonics as before then the same mechanical pitch will result in

$$y_2 = p_2 / p_1 \cdot y_1 = ... = N \cdot 180^0$$

electrical pitch for the higher pole number.

If N is even the electric angle for the higher pole number is always

$$y_2 = N \cdot 180^0 = ... = 360^0$$

It means that there will be no fundamental wave on $2p_2$ at all.

If N is odd, however, the electric angle for the higher pole number is always

$$y_2 = N \cdot 180^0 = ... = 180^0 \text{, full pitch}$$

This is the reason why the machine will work as a standard one on both pole numbers, but always with full pitch. No further

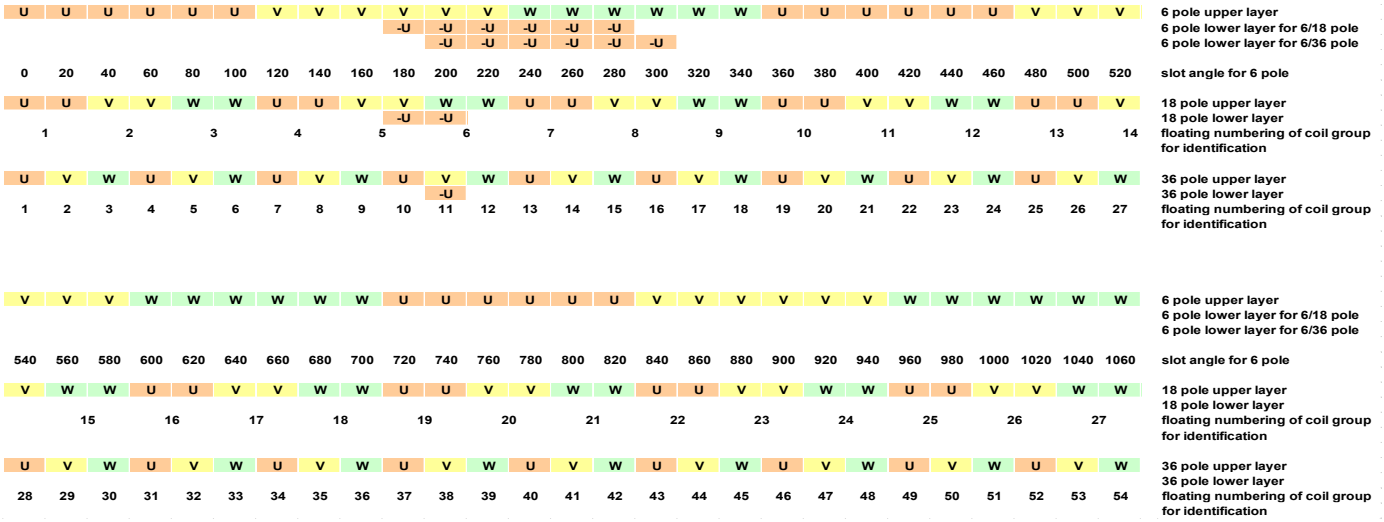


Figure 7. Slot distribution pattern for 1:3 as well as for 1:6 pole ratio

scientific considerations are required and the designer may apply basically his standard software (but for 120° phase belt).

However, further investigation shows that in such cases (N=3, N=6 etc.), phase asymmetry will always appear again, on the lower pole number only, the nature of which is the same as in *5.1*, but the magnitude is much higher. It is even so high that a solution must be found to eliminate it.

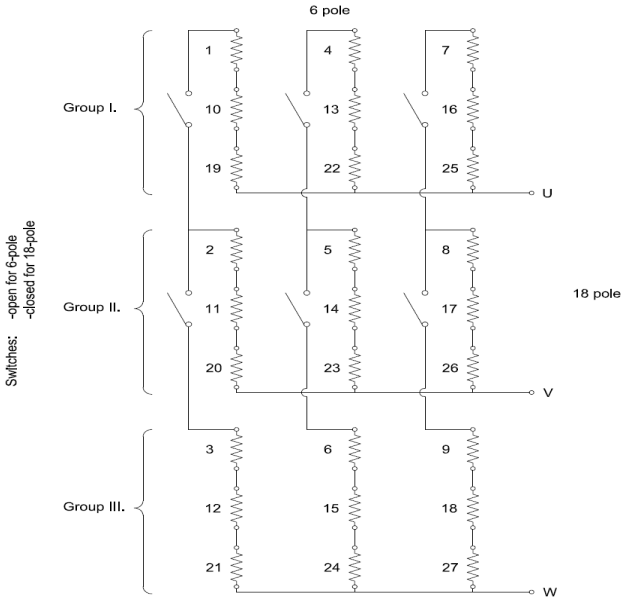


Figure 8. Pole changing connection pattern for 1:3 pole ratio

6.2. Specific Considerations for the Lower Pole Number

The lower pole number on Figure 7. will be discussed again. Comparing it with Figure 8.. it comes clear that Group I. contains always the "beginning part" of the pole, Group II. the "middle part", Group III. the "last part".

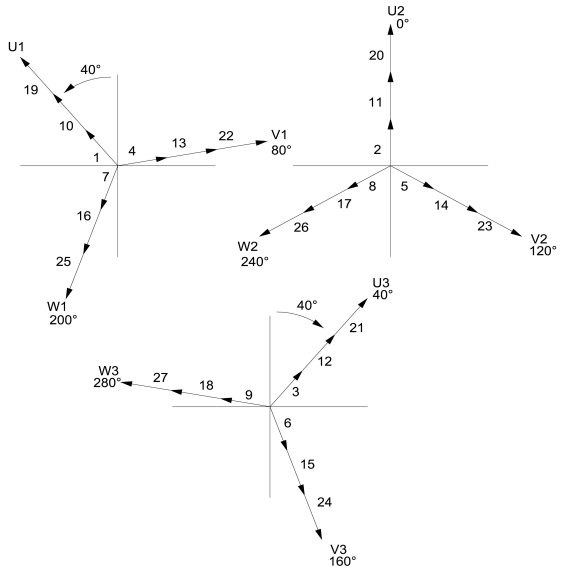


Figure 9. Voltage vector pattern for 1:3 pole ratio

Supplying the – parallel connected – Group I, Group II and Group III. with the same 3-phase system, then considerable balancing currents will be generated by that always 40° phase imbalance. Although there is wide ratio here, the approach of calculation method given in Appendix may be applied. Nevertheless, no need for any further calculation (neither that of extreme difference in the currents nor additional torque ripples in the motor caused by local air gap induction variation) to see that it exceeds any tolerance range.

Instead of looking for solutions in the machine (losing its simplicity) a solution in the supply is more recommended this time.

Group I. shall be supplied
by a 3-phase system -40° – 80° – 200°
Group II. shall be supplied
by a 3-phase system 0° – 120° – 240°
Group III. shall be supplied
by a 3-phase system 40° – 160° – 280°

Just a clear 9-phase supply system has been reached in this way.

Consequently, the only possibility for avoiding balancing currents in pole numbers 2p – 3x2p or 2p – 6x2p, etc. is to have nine terminals for the lower pole number.

Such solution was already proposed by [9], but through a completely different way of consideration and actual connection arrangement, and then continued in [10] and [11] but as a result of a particular consideration, only for 1:3 ratio.

For creating 9-phase system, frequency converters may be used. The converter is connected into 3-parallel (not unusual for high power) and the regulation of each parallel branch is shifted by 40° from each other, see Figure 10.

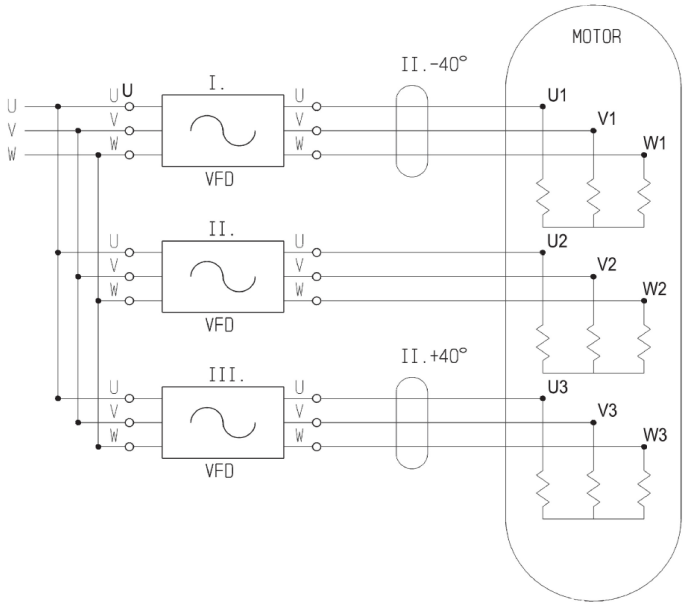


Figure 10. Proposal for supply of the stator winding on the lower pole number

As a consequence, no further derating other than the inherently k1 = 0.866 is necessary to apply in case of 1:3 (1:9 etc.) pole combination.

Such shifting of 40° is not necessary for the higher pole number, which may thus preserve its 3-phase winding – 3-phase supply – 3-terminal. When changing the supply to the higher pole number, the nine terminals of the lower pole number (always 3-3-3 of the same phase) must be short circuited by a separate switch.

The same energy supply concept shall be applied for the lower pole number in case of 6-36 pole, 6-54 pole etc. that means for 1:6, 1:9, etc. pole ratio as well.

Mechanical pitch corresponds always to that of the lower pole number. This mechanical pitch (endwinding construction) is very far from the usual ones at standard high pole motors. That is the only fact to be taken into consideration when electromagnetic calculation is being performed by standard calculation software.

6.3. Further extension of the approach: 1:4, 1:5, 1:7… pole ratio

A motor winding supplied by the usual 3-phase network has been investigated so far. Therefore, some limitations have appeared, such as 1:4, 1:5, 1:7 etc ratios have not been possible. If, however, this approach is exceeded, and the number of the phases is extended, as a very new approach, those missing pole ratios became possible immediately and very easily. 1:4 pole ratio is possible to achieve by 4-phase supply and 4//Y / 4//Y connection, 1:5 pole ratio by 5-phase supply system and 5//Y / 5//Y connection, 1:7 pole ratio by 7-phase supply system and 7//Y / 7//Y connection and so on.

As an example, the slot distribution pattern as well as connection pattern for 5 phase 5//Y / 5//Y connection is given on Figure 11. and Figure 12., resp. just for demonstration (for which the slot distribution pattern is only its beginning part, for the sake of good visibility and readability only). It is to be continued for N phase N//Y / N//Y connection in the same way.

Some conclusions found before still valid. Supply concept on the lower pole number must be applied similarly to Figure 10. in order to avoid those extreme balancing currents.

As a consequence, the number of the terminals will increase according to the number of the phases, also because NxN phase supply is necessary for the lower pole number.

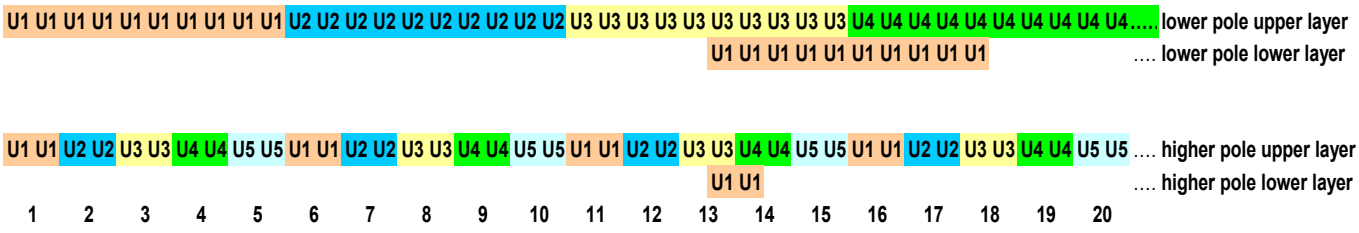


Figure 11. Slot distribution pattern for 1:5 pole ratio

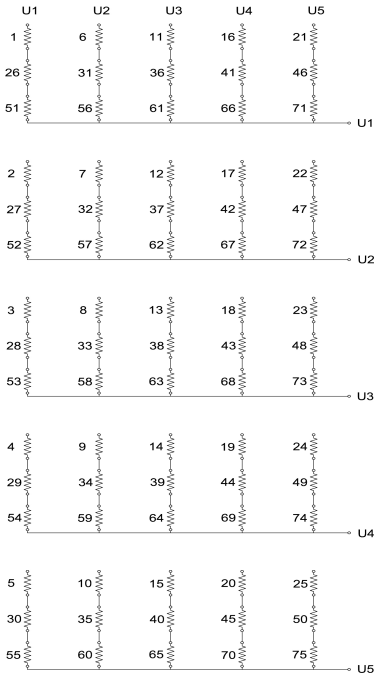


Figure 12. Connection pattern for 1:5 pole ratio (5 phase supply)

Further, what found before for the case if N is even or odd, is still valid. If N is even, like 1:4, even harmonics will appear. If N is odd like 1:5, 1:7 etc. the result again is absolute ideal, no even harmonic will appear on either pole number, therefore the motor's behavior on both pole numbers is identical to the corresponding single-speed motor. Derating factor of k1=0.866 is still valid.

Due to the high number of parallel connection – high number of phases is also a certain kind of parallel connection – the method is inherently suitable for high power.

6.4. Rotor slot numbers

The literature gives little support for how to choose the rotor slot numbers.

- q number of slots per pole per phase must not be an integer (for avoiding synchronous harmonic torque at zero speed)
- number of rotor slot number per 2p must be an integer (for avoiding imbalanced magnetic pull).

Considering these for both the higher as well as the lower pole number a clear picture can be obtained, preferable on a usual spreadsheet.

The result is the following:

- there is no common slot number on the sheets because
- each slot number proposed acc. to 2nd requirement for higher pole number are forbidden slot numbers acc. to 1st requirement for the lower pole number. The reason is that each gives an integer q for the latter.
- no any on the (long) list of proposed slot numbers of the lower pole number can be found on the (shorter) list of the higher pole number.

It is clear in this way that it is impossible to find a rotor slot number which would fulfill even those minimum requirements above at the same time.

What in this situation can be preferred is the lower pole number – because of the higher speed – unless the generally higher air gap induction on the higher pole number seems to make more problems.

6.5. Supply voltage strategy

It was obvious in the chapter *5*. for close ratio that the supply voltage for both pole numbers is the same. It is not necessarily the case, however, here. Therefore, it may be discussed shortly what supply strategy may be possible when switching from the lower pole number to the higher pole number and what the consequences are.

The formulas already found in chapter *5.3.6.* will be applied. The case for N = odd will be discussed only just for more compact formulas because here the pitch is identical (full) pitch on both pole numbers therefore: $\xi_{1p1} = \xi_{1p2}$ and the consideration can focus on the important phenomena.

The logical target – when determining the power - will always be to obtain identical maximum torque expressed in p.u. on both pole numbers (M_{max} /M_{rated} remains). Therefore, the ratio of the maximum torque absolute values will be expressed as being proportional with the square of the voltage.

6.5.1. Same voltage

$$U_{p2} = U_{p1}$$

$$M_{\max p2} = \left(U_{p1}/U_{p2}\right)^2 * p_2/p_1 * M_{\max p1} = ... = N * M_{\max p1} \quad (14)$$

$$P_{p2} = M_{ratedp2} * \omega_{p2} = N * M_{ratedp1} * \omega_{p1}/N = P_{p1}$$

$$I_{p2} = P_{p2}/U_{p2} = ... = I_{p1}$$

$$B_{p2} = N * B_{p1} \quad (!)$$

The operation corresponds to "constant power".

The result is obviously identical to that obtained in *5.3.6* for close ratio.

6.5.2. Voltage reduced by $\sqrt{N}$

$$U_{p2} = U_{p1}/\sqrt{N}$$

$$M_{\max p2} = ... = M_{\max p1}/N * N = M_{\max p1} \quad (15)$$

$$P_{p2} = M_{ratedp2} * \omega_{p2} = M_{ratedp1} * \omega_{p1}/N = P_{p1}/N$$

$$I_{p2} = P/U = P_{p1}/N/U_{p1}/\sqrt{N} = I_{p1}/\sqrt{N}$$

$$B_{p2} = N/\sqrt{N} * B_{p1} = \sqrt{N} B_{p1}$$

The operation corresponds to “constant torque”.

6.5.3. Voltage reduced by N

$$U_{p2} = U_{p1}/N$$

$$M_{\max p2} = \ldots = M_{\max p1}/N \qquad (16)$$

$$P_{p2} = M * \omega = M_{ratedp1}/N * \omega_{p1} \quad /N = P_{p1}/N^2$$

$$I_{p2} = \ldots = 1/N^2 * 1/N * I_{p1} = I_{p1}/N$$

$$B_{p2} = B_{p1}$$

The operation could be named as “constant stress of motor”.

It shall be remarked that all the above formulas regarding the current are valid only if the (eventual) change of η and cos φ is out of consideration.

It shall be remarked also that - when connected to p2 - the induction in the iron core will be changed in an unusual way:

B_{tooth} changes proportional to B_{airgap} ;

B_{yoke} changes proportional to B_{airgap} / N .

The right strategy may be chosen acc. to the torque requirements of the drive.

7. Conclusion

In this paper, a complete analysis of 3//Y / 3//Y pole changing method has been provided through a unique approach, including application, characteristics and design aspects in a comprehensive way, especially for high power motors.

The main finding of the analysis is that the method is suitable for any pole combination, with no limitation for both close ratio as well as for wide ratio pole changing windings.

It has been proven for close ratio that the method is well applicable without balancing measures for practically the whole range required by the industry. The limits are given and are supported by calculations.

Moreover, as a fundamental novelty, a solution for a theoretically infinite range of 1:N pole change ratio winding connection has been given (N phase N//Y / N//Y connection), with no need for further scientific considerations, that means it is suitable for a direct industrial use. The designer may use his standard calculation software.

A new, complete drive system for main marine propulsion has been proposed, with the pole changing motor with 1:N ratio as the core element.

8. Appendix

8.1. Approximate calculation of balancing currents discussed in connection of Figure 4.

The phenomenon was simple mentioned already in [6] but with little explanation, no calculation.

Because the phenomenon arises in no-load as well, even, the phenomenon is in connection with the main flux and the induced voltage by the main flux, the simplified equivalent circuit known from the theory may be applied.

This simplified equivalent circuit shall be applied for each parallel branch see Figure 13.. The voltage of the branches is the same: U, the supply voltage. The currents, however, are obviously different: they will be expressed by the sum of the common, resulting (magnetizing) current and the equalizing current I + ΔI. Regarding reactance and the induced voltage, not only the value but its direction must be taken into consideration in the meaning that the phase of the induced voltage will be different in different winding parts acc. to their position: that is why the reactance representing those “last 2 poles” are multiplied by $e^{j\alpha}$, *1*, $e^{-j\alpha}$ resp.

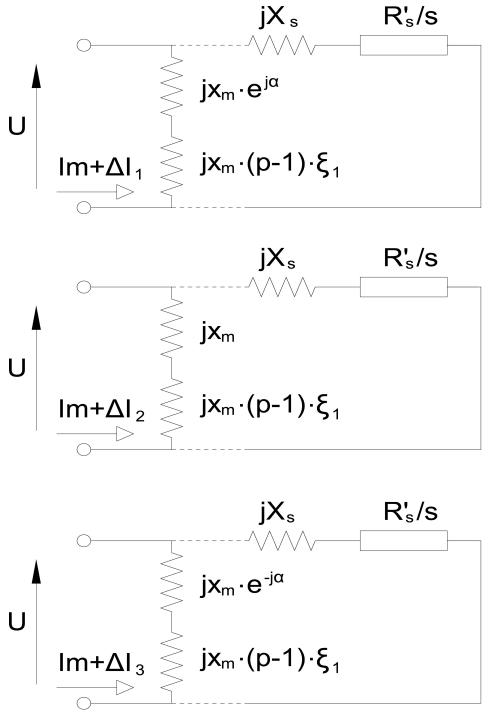


Figure 13. Simplified equivalent circuit

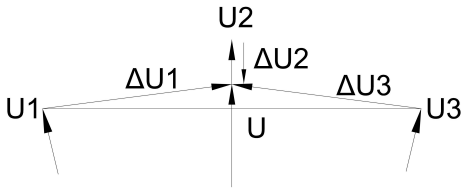


Figure 14a. Voltage vector diagram

Vector diagrams of the voltages and of the currents for no-load and load – based on Figure 4. - are shown on Figure 14a. and Figure 14b. The latter one is not in true scale.

The equations for the 3 parallel branches will be:

$$(I_m + \Delta I_1) * jX_m((p-1) * \xi_1 + e^{j\alpha}) = U$$

$$(I_m + \Delta I_2) * jX_m((p-1) * \xi_1 + 1) = U \qquad (17)$$

$$(I_m + \Delta I_3) * jX_m\left((p-1) * \xi_1 + e^{j-\alpha}\right) = U$$

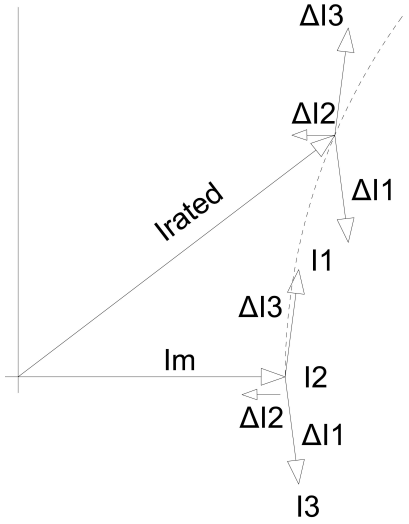


Figure 14b. Resulting equalizing currents

where

I_m - resulting, equivalent magnetizing current of one parallel branch

X_m - magnetizing reactance per pole per branch

$p-1$ - takes into account that the "first 6 poles" are full poles $e^{j\alpha}$;1; $e^{-\alpha}$

U - supply voltage.

Because the sum of the equalizing currents is per definition zero, if all three equations are added, it yields

$$I_m = \frac{U}{jX_m} \frac{1}{\left((p-1)\xi_1 + (2\cos\alpha + 1)/3\right)} \qquad (18)$$

Actually, although I_m is given in this way, the main target is to know ΔI_n compared to I_m as the imbalance in p.u.:

$$\Delta I_1 = I_m \frac{(2\cos\alpha + 1)/3 - e^{j\alpha}}{(p-1)\xi_1 + e^{j\alpha}}$$

$$\Delta I_2 = I_m \frac{(2\cos\alpha + 1)/3 - 1}{(p-1)\xi_1 + 1} \qquad (19)$$

$$\Delta I_3 = I_m \frac{(2\cos\alpha + 1)/3 - e^{-j\alpha}}{(p-1)\xi_1 + e^{-j\alpha}}$$

Magnitudes of the equalizing currents are (for 8 poles) approximately:

$\Delta I_1 \approx \Delta I_3 \approx 0.19 I_m$

$\Delta I_2 \approx 0.05 I_m$

The phase angle is also interesting in case of further considerations because e.g. ΔI_2 results in local decrease while ΔI_1 and . ΔI_3 do so in local (half so much) increase of airgap induction which causes torque ripples and local radial force fluctuation.

Above equations are valid for 8, 14, 20 etc. poles. In case of 10, 16, 22 etc. poles

$(p-1)$ shall be substituted by $(p-2)$ and

$e^{j\alpha}$, *1*, $e^{-j\alpha}$ shall be substituted by $2e^{j\alpha}$, *2*, $2e^{-j\alpha}$ resp.

Conflict of Interest

The author declares no conflict of interest.

References

[1] G. Kovács, "A 3Y / 3Y Winding for high Power Asynchronous Motors", World Academy of Science, Engineering and Technology International Journal of Electrical and Computer Engineering Vol:12, No:1, 2018. p. 1367-1374

[2] Lengyel Zoltán, "Korszerű kéttekercselésű motorok", in Hungarian. Ganz Közlemények, Proceedings of Ganz Electric Works 1964. p. 31 – 36.

[3] G. H. Rawcliff, B. V. Jayawant, "The development of a new 3: 1 pole-changing motor". Proceedings of the IEE – Part A Power Engineering 1956. p. 306 – 316.

[4] T. H. Barton, O.I. Butler, H. Sterling, "The theory and Characteristics of the new 3:1 pole changing induction motor", Proceedings of the IEE – Part A: Power Engineering 1956 p.285 – 294.

[5] K.C. Rajaraman, "Design criteria for pole-changing windings", Proceedings of the Institution of Electrical Engineers (Volume: 124, Issue: 9, September 1977) 775 – 783, https://doi/org/10.1049/piee.1977.0171

[6] A.R.W. Broadway ; K.S. Ismail , "Phase modulated 3-phase pole changing windings", IEE Proceedings B - Electric Power Applications 1986 p. https://doi.org/10.1049/ip-b.1986.0011

[7] P. Chidambaram ; M. Subbiah ; M.R. Krishnamurthy, "Developments in Pole Changing Windings using Star/Star-Delta Switching", IEEE Transactions on Energy Conversion (Volume: EC-1, Issue: 4, Dec. 1986) p. 135 – 141 https://doi/org/10.1109/TEC.1986.4765787

[8] K.C. Rajaraman, W. Ranjith Perera, G. Subramaniam, "New wide-ratio pole-changing motors", IEE Proceedings B - Electric Power Applications (Volume: 134, Issue: 4, July 1987) p. 211 - 214 https://doi.org/ 10.1049/ip-b.1987.0036

[9] J.W. Kelly ; E.G. Strangas ; J.M. Miller, "Control of a continuously operated pole-changing induction machine", IEEE International Conference on Electric Machines and Drives. 2003 IEMDC'03. https://doi.org/10.1109/IEMDC.2003.1211265

[10] Gautam, J. Ojo, " Variable speed multiphase induction machine using pole phase modulation principle", IECON 2012 – 38th Annual Conference on IEEE Industrial Electronics Society 2012. https://doi/org/10.1109/IECON.2012.6389310

[11] Umesh B S, Kiran Kumar N, Sivakumar K, "Performance improvement of a nine phase pole phase modulated induction motor drive". IEEE International Conference on Industrial Technology, 2015 p.812-817. https://10.1109/ICIT.2015.7125198

[12] Baoming Ge ; Dongsen Sun ; Weiliang Wu ; Fang Zheng Peng , "Winding Design, Modeling, and Control for Pole-Phase Modulation Induction Motors" IEEE Transactions on Magnetics (Volume: 49, Issue: 2, Feb. 2013) p. 898 - 911 https://doi/org/10.1109/TMAG.2012.2208652

[13] Dawei Meng, Qian Wang, Research on Balanced Winding Used in 3Y/3Y pole changing Motor, - International Journal of Automation and Power Engineering (IJAPE) Volume 2 Issue 4, May 2013

[14] H. Liu, J. Wang, Z. Zhang, Performance analysis of variable speed multiphase induction motor with pole phase modulation, Archives of Electrical Engineering, Vol. 65(3) 2016. p. 425-436

[15] W. Nürnberg, "Die Asynchronmaschine", in German, Springer Verlag 1952. p.299.

[16] Liska József, "Villamos gépek, IV. Aszinkron gépek", In Hungarian, Tankönyvkiadó Budapest 1960. p. 126.

[17] Bedrich Heller, Václáv Hamata, "Harmonic Field Effects in Induction Machines", ELSEVIER Scientific Publishing Company Amsterdam, Oxford, New York 1977

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 243-246 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Closed Approach of a Decoder Mobile for the 406 Mhz Distress Beacon

Billel Ali Srihen[1,*], Jean-Paul Yonnet[2], Malek Benslama[1]

[1]*Laboratory of Electromagnetism and Telecommunications, University of Brothers Mentouri Constantine, 25000 Constantine, Algeria.*

[2]*Laboratory of Electrical Engineering, University Joseph Fourier Grenoble, 38000 Grenoble, France.*

A R T I C L E I N F O

Article history:
Received: 16 August, 2018
Accepted: 28 October, 2018
Online: 10 November, 2018

Keywords:
Frame Decoder
COSPAS SARSAT
Decoder Mobile
GPS coordinates
406 MHz Distress Beacon

A B S T R A C T

This article presents the design and realization of decoder mobile for distress beacon 406 MHz. The use of the sparse Fourier transforms in the processing of distress beacon signals. The characteristics of these decoders are mobility, able to decode all the frequencies of the beacons. It locates with great precision the location of the emitting emergency beacon, which is equipped with GPS, indicates the actual distance from which the alerts emanate. This compressed detection can give us a more detailed form of signals, so new results are given in the detection, the performance of the results is presented through experience.

1. Introduction

Canada, France, the Soviet Union and the United States are the first countries to sponsor the COSPAS SARSAT system at its inception which is a permanent international humanitarian satellite search and rescue system, detecting and locating beacons of emergency installed in ships, aircraft, and individuals. The program was subsequently joined by many other countries today (total forty countries and two organizations in 2016) [1] [2]. Regarding distress in the three cases cited, the finding the position of the disaster, and the necessary time of rescue and effectiveness depend on the reliability of the communication between the LUTs and the satellites [3] [4]. In order to provide semi-realistic alerts and location information from GEOSAR, the second-generation 406 MHz beacon is used to allow the location data to be encoded in the transmitted 406 MHz message [5]. The importance of this research is to reduce the time spent in research during the rescue process through the use of decoding screens, which in turn contain information inside the frame (GPS coordinates and the distance between the mobile decoder and markup), which enables us to identify the location and the implementation of the rescue process quickly.

2. Cospas Sarsat Concept

Figure1 presents a basic structure of the COSPAS SARSAT system [1] [2]. In the most disasters cases, the emergency beacons are activated when human lives are threatened. These alerts are received by the satellites and are retransmitted to the terrestrial stations distributed throughout the world "Local User Terminals" (LUT). Alerts are routed to the country's Mission Control Center (MCC), which operates the satellite terrestrial stations from which LUT is located. Routing alerts include the location of the calculated beacon to the LUT received by one of the orbiting low-Earth (LEO) satellites.

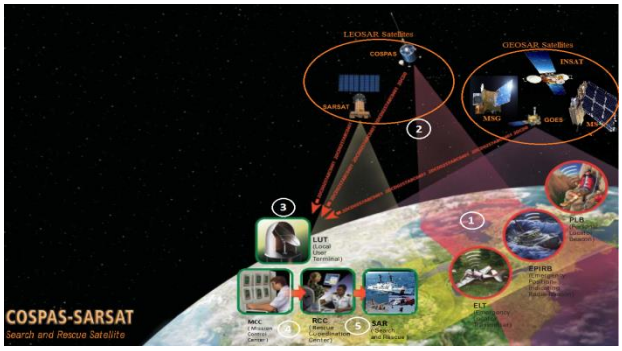


Figure 1. COSPAS SARSAT system [1]

The rescue operation is a carefully coordinated phase. Initially, the alerts are verified (based on the Doppler effect) after the confirmation the information is sent according to the country of the registration or to any rescue coordination center (RCC), which is equipped with emergency search and rescue correspondence [1] [2]. SAR satellites can also receive low distress signals from beacons.

After 2009, all beacons transmit on 406 MHz [6]. The universe is so complicated; the human being is always thirsty for discoveries enabling him to discover it, which exposes him to a danger, which

*Corresponding Author: Billel Ali Srihen, E-mail: bil_rst@yahoo.fr

has led through research and practice, the creation and realization of three types of emergency beacons currently in service [1] [2]:

a. Personal Locator Beacons (PLBs) for individuals,

b. Emergency Position Indicating Radio Beacons (EPIRBs) for maritime applications, and

c. Emergency Locator Transmitters (ELTs) for aviation applications.

3. Frequency beacons cospas-sarsat 406 mhz

In January 2010, the 406.040 MHz frequency began to be used. Where the principles frequencies are 406.049 MHz and 406.052 MHz [6].

Where, the band 4 0 6.0 – 4 0 6.1 MHz has been set aside by the International Telecommunication Union (ITU) for the low power satellite emergency position Currently, Cospas Sarsat distress beacons transmit to 406.025,406.028, 406.037, and 406.040 MHz.

4. Specification and Design

4.1. Specification

Specification of the second Generation 406 MHz SAR beacon and specification of ELT is defined in COSPAS-SARSAT technical document [5] [7].

4.2. Functional block diagram of decoder

Figure 2 shows the Mechanism functioning of a 406 MHz decoder. The main objective of the decoder is to identify the system, which that retransmits, determine when the reception took place, and then retransmit it again. The information to retransmit the weft shall be received as well as the temporal information. All the information's are displayed on one single screen with 4 lines of 20 characters. To do all this, the information is intensified.

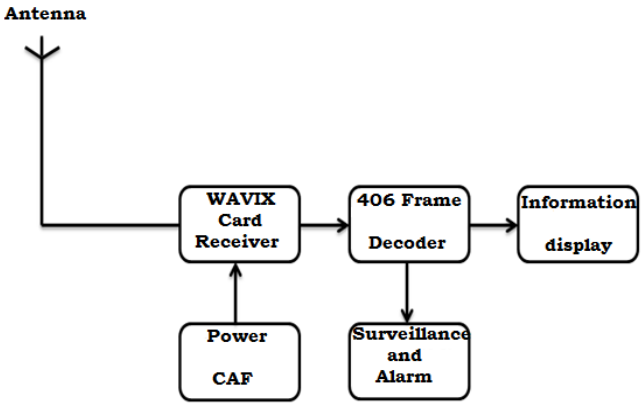


Figure 2. Functional schema of a 406 MHz decoder

5. Implementation of the Decoder

Figure 3 and 4 presents the Frame Decoder and complete system respectively. These two figures show: on the left side the jack 2.5 mm for the GPS connection, one on-off switch in the middle and the power jack which also serves as the battery charge point, and the jack 3.5 mm to connect the output receiver on the frequency of the beacon.



Figure 3. Frame decoder with display homepage.

On the right side, LED is used to monitor the operation and two push buttons allow navigating in memory (Figure 9). Since the system permanently records new message data, it is preferable to disconnect the entry when you want to see the pages in memory.

The complete mobile decoder after improvement of the old decoder. The system "receiver decoder" is now operational tobe with the permanent listening of the beacons of distress is shown in Figure 4 [10]. Its sensitivity is very good. As soon as a beacon passes in emission, whatever its frequency, the receiver is fixed automatically on the real frequency of the beacon and the decoder displays the contained information in the screen. The sound of the buzzer is then made hear during one second.

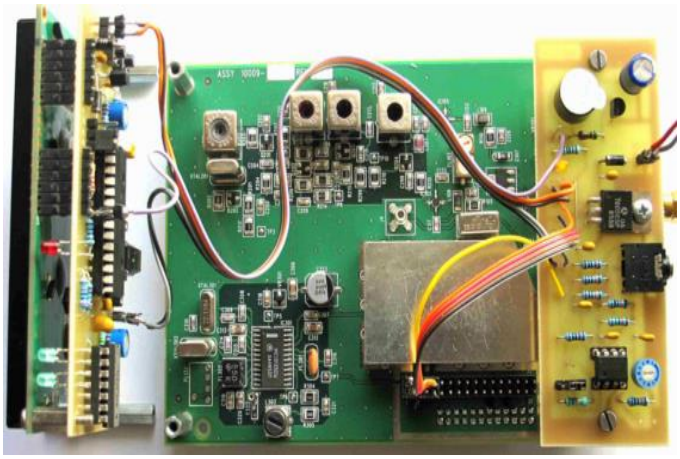


Figure 4. The complete system for receiving and listening to 406 MHz beacons

6. Experiment Results

6.1. Performance of the ELT Beacon

The 406.040 MHz ELT beacon signal is experiment results. The frequency meter and Spectrum Analyzer indicates the frequency is 406.040 MHz [10], it has the same spectrum of an EPIRB Beacon [11].

6.2. Position of the Beacon

Figure 5 corresponds to GPS position transmitted by the beacon. The frame is containing the position displayed on the fourth line. Where, all information is present except the time of receipt.



Figure 5. The GPS position transmitted by the beacon

The experience of the GPS receiver. Figure 6 shows Longitude Est. 02.0955^0 and North Latitude 45.4700^0 shows position in Google Earth. This indicates the point indicated in the figure this experiment was carried out.



Figure 6. Position Plot in Google Earth

6.3. Functioning of frame decoder

Output of the receiver listening to the beacon

The signals emitted by the beacon are a series of peaks, alternately positive and negative. These peaks are 1.25 ms or 2.5 ms. they correspond to the phase jumps of the PSK modulation as shown in Figure 7 and Figure 8 . In fact these peaks crossepond to the rising brows and the descending brows of the crenels that control the PSK modulator.

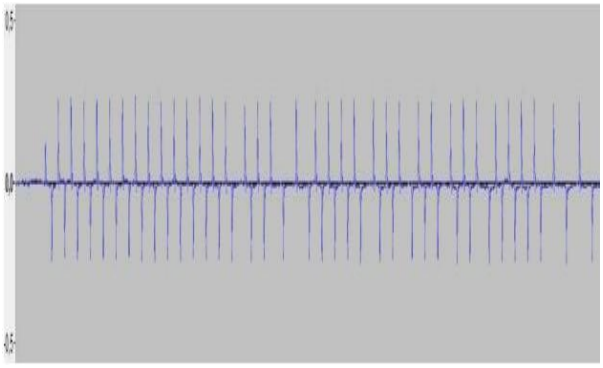


Figure 7. The frames emitted by the beacon

Figure 8 shows the signals at the output emitted by the Beacon; the frame information is more scope by battlements but by peaks. With this type of output by peaks, the frame generator performs an excellent simulation of the chain "Beacon 406 + receiver"

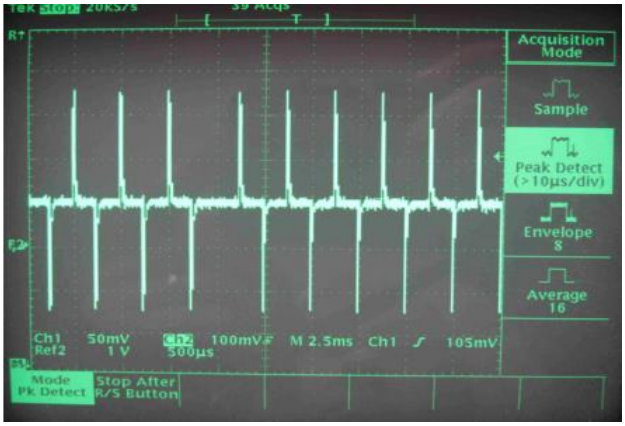


Figure 8. The signals at the output in the form of alternating peaks

Display of the distance between mobile decoder and the beacon

The calculation and display of the distance are fully automatic. As soon as the decoder knows its position and that of the beacon, it displays the distance and heading in place of the identification of the beacon on the second part of the third line. Figure 9, and 10 shows the distance display: it is 10:09 Local time, and heard beacon is a 49.9 Km north west.



Figure 9. Display of distance and heading

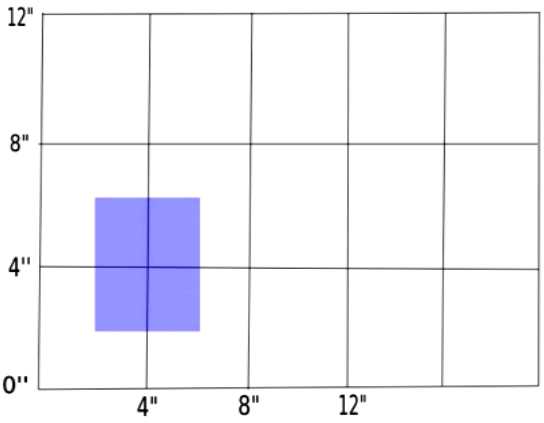


Figure 10. The position of the beacon is discretized at no 4 arc seconds, which leads to a position of uncertainty 123 m North and 87 m East

7. Conclusion

This work concerns the disaster beacon, even at great distances, and when the signals are weak, it is quite easy to recognize the signal of the 406 MHz beacon in the noise.

By coupling the receiver with a 406 MHz fram decoder, a beacon receiving station is obtained, which can operate as a permanent listener, and it will automatically decode the transmitted frame, so that our receiver-decoder is operational to be permanently listened to The reception of a signal on any frequency, the receiver automatically stabilizes on the frequency of the beacon and our decoder displays the information contained in the frame (GPS coordinates and distance between Mobile decoder and beacon), which will allow us to quickly locate the place, and facilitate the intervention of rescues, consequently, reduce the search time. According the new development in monitoring beacon as showing by [13], we can expect more interesting results in the using of disaster beacon.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was done in the framework of a partnership between the University of Brothers Mentouri Constantine Algeria and the University Joseph Fourier Grenoble 1 France. Under the direction of CNRS Research Director emeritus Jean-Paul Yonnet.

References

[1] http://www.cospas-sarsat.org, 30.05.2018.
[2] http://www.sarsat.noaa.gov, 30.05.2018.
[3] C. W. Scales and R. Swanson, “Air and Sea Rescue via Satellite Systems”, IEEE Spectrum (ISSN 0018 – 9235), vol 21, pp. 48–52, March 1984. https://doi.org/10.1109/MSPEC.1984.6370206
[4] I. W. Taylor and M. O. Vigneault, “A neural network application to search and rescue satellite aided tracking (SARSAT), In Proceedings of the Symposium”, Workshop on Applications of Experts Systems in DND, pp. 189-201, Royal Military Coll. Of Canada, 1992.
[5] J. V. King ‘‘ New Developments in the COSPAS-SARSAT Satellite System for Search and Rescue’’, 55th International Astronautical Congress 2004 https://doi.org/10.2514/6.IAC-04-M.4.07
[6] Specification for COSPAS – SARSAT 406 MHz Distress Beacons, C/S T.001 Issue 3 – Revision14, October 2013
[7] COSPAS-SARSAT 406 MHz Frequency Management Plan, C/S T.012Issue 1 – Revision 9, October 2013.
[8] COSPAS-SARSAT 406 MHz Distress Beacon Type Approval Standard, C/S T.007 Issue 4 –Revision8, October 2013.
[9] COSPAS-SARSAT Guidelines on 406 MHz Beacon Coding, Registration and Type Approval, C/S G.005, Issue 2 - Revision 5, October 2010.
[10] B ALI SRIHEN, J.P Yonnet, and M BENSLAMA, ‘‘Design and realization an ELT beacon and decoders of frames 406 MHz’’. International Conference (IEEE) on Engineering & MIS (ICEMIS), 8-10 May 2017, Monastir, Tunisia https://doi.org/10.1109/ICEMIS.2017.8272967
[11] I. Joo, J.H. Lee, Y.M. Lee, C.S. Sin, S.U Lee, and J Kim, ‘‘Development and Performance Analysis of The Second Generation 406 MHz EPIRB ’’ International Conference (IEEE) 4th Advanced Satellite Mobile Systems, 26-28 Aug. 2008, Bologna, Italy https://doi.org/10.1109/ASMS.2008.64
[12] S. STATLER ‘‘ Beacon technologies’’ Apress 2016; p 21; p 82; San Diego; California, USA

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 248-257 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Ubiquitous and Configurable Wrist-Worn Sensor Node in Hazardous Gases Detection

Mostafa Haghi[1,*], Kerstin Thurow[1], Regina Stoll[2]

[1]*Center for Life Science Automation (celisca), Rostock University, Rostock 18119, Germany*

[2]*Institute of Preventive Medicine (IPM), Rostock University, Rostock 18055, Germany*

ARTICLE INFO

Article history:
Received: 27 September, 2018
Accepted: 04 October, 2018
Online: 08 October, 2018

Keywords:
Wearable sensor
Ambient monitoring
Hazardous gases
Configurable device
Prolonged monitoring

ABSTRACT

The monitoring of environmental parameters is a critical topic in occupational medicine. With advances in technology, during the recent years, the wearable sensors are widely used to evaluate the ambient parameters. In this paper a compact, small, light-weighted, and configurable sensor node in three physical layers for prolonged chemical ambient monitoring is described. The device is designed and demonstrated by using MLMS (Multi-Layer Multi-Sensor) concept. The device application is not only limited to hazardous gas (CO & NO_2) monitoring but physical ambient elements (air pressure, air humidity and temperature) and motion tracking are tracked, too. The device consist of integrated (ambient physical and motion tracking) and add-on sensors (hazardous gases) and enables centralized data processing. The integrated sensors are distributed on the middle physical layer on both sides (host platform). The host platform also includes micro controller and antenna for data transmission to the smartphone. The gas sensor along with the gas sensor driver board form the gas sensor node which for easy gas exposure is located at the top of the device and stuck at the top of host platform. The gas sensor driver is designed universally to be compatible with 3-lead gas sensors with low power consumption. This layer is replaceable and can be substituted with one more target gas (e.g. NO_2 instead of CO). The sensor configuration and initialization is performed via sending commands from the smartphone to the device by the wearer. In addition, the device is equipped with the notification system (beeper and vibrating motor) for early user warning in an abnormal detected. The notification driver layer is located at the bottom of the proposed device. All physical layers are stuck on top of each other appropriately through a board to board connector. The collected data from integrated and add-on sensors are transmitted in various time interval to the smartphone for screening and logging. Filter design and calibration process are also briefly described in this work.

1. Introduction

1.1. Ambient air monitoring and wearable devices

Measuring hazardous and toxic gases such as nitrogen dioxide (NO_2) and carbon monoxide (CO) in the environment (indoor/outdoor) and also meteorological parameters (atmospheric temperature, humidity and pressure) is usually expensive. Increasing number of elderly, and costly healthcare monitoring systems in hospitals from one side and high degree of penetration of wearable devices (applied to wide range of applications) in particular healthcare and tele-monitoring on the other side, attract much attention as a potential solution. This paper is an extended version of presented paper in [1]. The work is extended in terms of filter design, calibration process and providing more details. Furthermore, the device is configurable in the new version; switching between different gas sensors is possible through sending commands from the smartphone to the device.

Information on air-quality in urban environments due to cost of equipment and space is typically measured only in limited number of sites [2]. In recent years, wearable devices and technologies are being popular and demand is increasing for the application in ambient air monitoring. Especially employees who are working in chemical, biological or medical laboratories can be exposed to gases which might affect the individual performance and health.

[*]Mostafa Haghi, Center for life science automation, F. Barnewitz-Str. 8, 18119 Rostock, Germany, +49-3814987806 & Mostafa.Haghi@celisca.de

The form factor (wearability), enabling technology, accuracy, low cost and prolonged ambient monitoring are some of the most important criteria that can push and increase the penetration and popularity of the wearable devices in the environmental parameter monitoring. The non-linear response of the chemical sensors in ambient air monitoring, is significantly affected by the environmental conditions and the applied heating pattern for the sensor initialization [3], especially in the regime of the on-demand sensing including off-time periods [4].

Electrochemical gas sensors due to the acceptable merit and criteria are the adequate candidate to be the major player in a wearable/portable gas measurement device. These merits include the very low power consumption, high sensitivity, repeatability and mostly the electrochemical gas sensors from the wide and different range of families, approximately have a linear response under controlled conditions [5]. The linear response may extensively be useful in calibration and recalibration. Although, the gas sensors are relatively costly in personalized wearable device application, but recently it is getting closer to an affordable range, especially in mass production. In some cases, (e.g. Spec sensors) the life time of sensors are expected between 5 and 10 years.

One of the major contribution of this work is multi-parameters measurement, while the solution is still compact, small and wearable. In fact, the effort is placed to draw the attention from the 2-D design to 3-D. Therefore, z axis also is added to the calculation of prototype implementation. In this new concept known as Multi-Layer Multi-Sensor (MLMs), each physical layer is dedicated a particular function and these layers are stuck on top of each other. This would shrink the x-y plane size and slightly extend the altitude [6]. This concept deploys board to board connectors to link the layers.

The second important features of this work is the continuous monitoring for several hours without need of battery charge. This is accomplished through careful electronics circuit design and hardware selection. Where, of the critical application of ambient monitoring is in personal healthcare monitoring, thus a longer observation of ambient parameters may protect the individual health status for longer. In other word, the larger data acquisition database is, the larger information may be provided for health status analysis. However, as the tool is consist of the notification system, if the air is unhealthy, the warning signal is generated for protection.

Creating a data base and data analyzing for the health status of the individual long-term monitoring has become an aim to enable an early notification of individual health status by analyzing the large amount of collected data. The prolonged monitoring is the function of power consumption. The power mostly is dissipated by the driver, sensors and data transmission. As a consequence, the sensor selection, driver electronics circuit design and frequently data transmission to the smartphone (time interval) and sensor sampling rates, must be carefully investigated. Once the total energy consumed by the device is estimated, the battery should be selected which mostly determines overall size and weight of the sensors and therefore user's convenience [7].

The gas sensor in this wearable is the most effective indicator in the final prototype in terms of form factor, power consumption and device performance. Conventional gas detectors including the most widely used (metal oxide), relatively small Pallisters require large power consumption (hundreds of milli-Watts to Watts) with slow response time (tens of seconds). The technical specifications for gas sensors, indeed for all sensors, are challenging, and typically [8]: small and light weight, low power (e.g. less than 200 mW), long life (e.g. 5 years), high selectivity (e.g. low cross sensitivities), range of operation -50 °C to +125 °C, range of operation between 5 and 95% relative humidity.

This paper is structured as follows: In section II some related research works are introduced. In part III, we concentrate on the proposed system architecture and the main components are described. In section IV, filter design, configurability and experimental results of a possible application scenario are presented to validate the efficiency of the proposed new multi-layer architecture design. Conclusion is summarized in section V.

1.2. Related works

With the great improvement in semi-conductor and wearable technology, the criteria for implementing a qualified device in the area of ambient monitoring in wearable devices are becoming more and more sophisticated to demonstrate. Multi-tasking operation, prolonged evaluation, flexibility and user friendliness are more crucial among others. In addition, flexibility in software and hardware extension for further development without starting from the scratch design, is a point that may not be forgotten. Future developments can upgrade the device with the latest technology development and move on the edge.

This section includes two general subsections: first the latest efforts by academic community is briefly described and after in the second subsection the available industrial tools in the marker are presented in this area of interest. The features, characteristics and most critical merits for each is addressed.

Crowd sensing is one of the major methodology in the ambient air screening, a sensor node by using this technique is presented by Oletic et al. in [9]. In this approach, the individuals are able to participate in the experiment by the infrastructure of a network of mobile sensors, to measure the factors. The device is a fixed structure sensor node with the constant and predetermined sensors which is powered on by a li-ion battery. The tool consists of two electrochemical gas sensors, temperature, and relative humidity and atmospheric pressure sensors. These sensors are sampled and data are transmitted to the smartphone with Bluetooth. In [10] Ritcher and et al. introduce a wearable environmental monitoring called Eco-Mini. This prototype is designed to investigate the feasibility of wearable device in ambient air monitoring in clinical studies. In this preliminary work, the wide range of toxic/hazardous gases (ozone, sulfur

dioxide, volatile organic compounds) are evaluated in a predefined mode. Furthermore, noise, humidity, temperature, and ambient light color balance are monitored. In addition, motion tracking is performed through 3-axis accelerometer. This prototype also is equipped with location (GPS). In spite of advantaged and a wide range of monitoring, this device is suffering from some features and critical requirements. This first notable issue is the fixed structure which does not give the flexibility to the user to select the sensor according to the needful. Furthermore, Eco-mini is connected to a smartphone and stores the sampled data locally through a SD card. The wearability mode of this tool is the next concern (waist-worn) which may lead to daily routing activity user interference. Sanfilippo et al. in [11] is using Cooking hacks platform to perform an investigation. In this work the physical parameters are supported to the vital signs (blood pressure and heart rate) sensors to screen the broader range of parameters which play role in health status. The whole structure is wire based. The different sensors are communicating to the platform to transmit the data, these data are sent to the smartphone through Bluetooth Low Energy (BLE). The last device from the academic community in the first subsection is a wrist-worn watch that published by Haghi and et al. [12], [13]. This watch is introduced for monitoring several physical parameters (air pressure, temperature and humidity) as well as toxic/hazardous gases (CO, NO_2). In addition, the motion tracking through 9 degree of freedom (DoF) by the gyroscope, accelerometer and magnetometer sensors is demonstrated. Tis device, deploy 3-lead electrochemical gas sensors to detect and evaluate the target gases. A notification through a beeper and vibrating motor are generated for user attention in abnormal status detection. In addition, the prototype utilizes BLE 4.1 to transmit the data to the smartphone in a predefined time interval.

In the second subsection which the attention is paid to the available industrial tools in ambient air monitoring, a brief description, advantages and disadvantages of each is outlined. In spite of costly devices in the market, these are mostly single task and parameter detector which are only produced for a particular gas detection. This makes these tools hardly affordable for the personal monitoring. However, this is not only obstacle. This gets critical when a user intend to detect several gases in different working places. Gas alert extreme device is introduced in [14]. This well-known device is an example of such devices. Gas alert is available in the market in a medium size and waist worn mode wearability. The device is already available for NO_2 and measure the concentration of up to 5 part per million (ppm). This range is extended due to the nature of CO up to 1600 ppm. It's equipped with alert notification and vibration motor as well. The disadvantages of Gas alert is lack of data storage in an external memory nor smartphone. The data are only observable through display and in real time.

In [15] Texas instrument (TI) introduces Single-Chip tool based on MSP430. This portable/ waist-worn is a single gas detector of TI and detects carbon monoxide (CO). Due to gas calibration, humidity and temperature sensors are integrated on board. Data storage in an external memory and real time data screening is of advantages in this costly TI product. This gas detector is provided in rectangular shape and is slightly larger than gas alert extreme.

The next presented industrial gas detector in this subsection is a circular single task product of TI. This evaluator is the CO gas detector [16]. This circular device with relatively high altitude that make it more portable than wearable, is capable of being connected to both Android and iOS. The gas detector has the capability of monitoring different hazardous gases by changing the sensor. Lack of display, external memory and notification system are the disadvantages. However, the device does not support any other physical parameters.

Jain et al. introduced a wearable environmental monitoring system (WEMS) [17]. In this platform OAQ for SO_2, NO_2 and O_3, in addition to UV radiation, humidity and temperature is evaluated. It is utilized a display for data monitoring and transmit the data through BLE to a smartphone. This WEMS is wrist worn. The platform is based on MSP430 TI dual processor with M4 and maximum rate of the frequency is 168 MHz WEMS consumes 6.5 mAh in standby mode and 98 mAh in watch display mode, this power consumption in sensing mode is 250 mAh.

In this work, the infrastructure approach based on the physical layers toward design and implementation of a personalized environmental parameters monitoring is introduced. To maintain the solution, small and wrist-worn the configurability through sending command from a smartphone and sensor node replaceability as the top layer is investigated. This features enable the approach to be more flexible to be applied in different applications according to the user decision.

2. Approach, Methodology and Layers' Description

Nowadays with excellent improvement in semiconductor technology, the size of sensors is shrinking continuously. Thus, the integration of numbers of sensors of different types in the same device becomes possible. Therefore, it is expected to reduce the dimension of the tools accordingly. This is one of the reason that the wearable sensors are coming to foreground and playing a more important role in different fields of applications (e.g. healthcare, smart city, smart home, agriculture and etc.).

The presented device (hereafter called Multi-Layer Multi-Sensor Environmental Monitoring-Gas (MLMS-EMG-3.2)) is working based on MLMS (Multi-Layer Multi-Sensor) concept [12]. MLMS is applied to create a wrist-worn device by shrinking the size of X-Y and extending the solution to Z vector (XYZ-42×42×25mm). In traditional approach more often the sensors and components are located side by side across the X-Y dimension, and length and width of the board is expanded. In the presented approach in this work, the X-Y plan plus Z are involved in development. To reach a wrist-worn in the design, the width and length of the device is not supposed to be expanded. In MLMS approach, each task is dedicated to a physical layer with its own sensor(s) and components. The signaling bridge between the

different layers is created through board to board connectors. In addition to this, this connector also stuck the layers on top of each other in a firm and reliable way. The gas sensor node, host platform and notification system are the three physical layers in this approach. However, the solution might be extended to further parameters measurements and more number of layers, while still being compact, small and light-weighted. Each layer specifications and features are described in the following sections.

In a large perspective MLMS-EMG-3.2 is working as a client and sends the collected data from the ambient physical and chemical sensors to a smartphone via Bluetooth low energy (BLE) for logging. The data for each parameter also are observed on the smartphone. In the following will realize that, the communication between smartphone and the wrist-worn is bidirectional. In the large and completed system, the data are transmitted through Wi-Fi to a server for a permanent storage and further analysis (Figure 1).

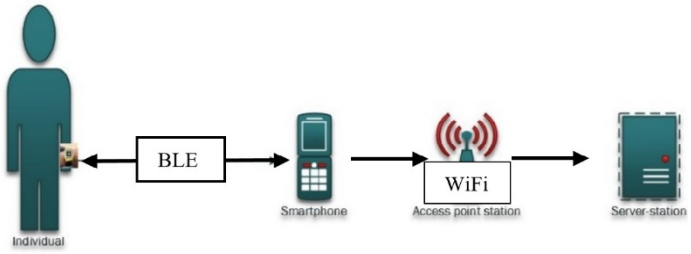


Figure 1 The general structure of whole system.

2.1. Physical Layers' Description

This approach consists of three physical layers, each dedicated to a special tasks. The layers are physically separated but are linked via board to board connectors to construct the whole device and to communicate to the micro controller and the host platform. Each layer is described as the follows:

- Middle Layer (host platform): host platform is the motherboard of MLMSEMG- 3.2. A careful host platform selection according to the requirements and criteria is mandatory to enable the integration of different sensors in this wrist worn device. The selected platform must be capable of expansion from both sides (z vector) physically. In addition, to support the number of sensors, a variety of communication protocols (ADC, SPI and IIC) is necessary. The platform should also enable the data transmission. Thus the iProtoxi host platform board based on nrf51822 (Nordic semiconductor) is the most adequate host platform for this prototype [18]. All three motion tracking sensors (gyroscope, magnetometer and accelerometer) and physical parameters measurement sensors (temperature, humidity and air pressure) are integrated on the board. These sensors are distributed on both sides of the board and allow the initialization through IIC communication. One of the extremely critical notable point is the future development of the prototype. The designers are intended to maintain the development extendable for more number of add-on sensors. This is possible with the two-side expansion in z axis through integrated board to board connectors. On the other hand due to limited number of pins on the board, the IIC communication is chosen to be applied to the all integrated sensors for ease of use. Even the gas sensor as an add-on sensor is partially use the IIC for driver initialization and programming.
- Top Layer (gas sensor node): gas sensor node is located at the top layer stuck on top of the host platform for ease of use (replaceability) and also free exposure to air. The gas sensor node is formed of two layers: gas sensor and gas sensor driver. A hardware independent, lower power consumption and flexible sensor node is demonstrated for efficient functionality. In general the gas sensor driver is compatible with 3-lead gas sensors (with some physical and size limitation), but each sensor has its own configuration, gain and filter design. Thus, these components (gain and filter) might be added externally by the designer according to the requirements. For ease of use, the gas sensor and the gas sensor driver are soldered together to form the gas sensor node. This sensor node is the layer at the top. Where the filter design and gain of driver board is the specific parameters for each gas sensor, the general architect and structure of the gas driver is the same for gas sensors but it is customized for in these specific parameters. As the conclusion, each gas sensor is coming with its driver. Furthermore, the soldered gas sensor node, protect the sensor from frequently touch by the user and possible damage.
- Bottom Layer (notification system): The warning system is a mean to reduce the potential risk. When the monitored parameters (here toxic/hazardous gas) are exceeding the thresholds, the notification system detects an abnormal status and is activated. The notification system is consist of a driver board that activate the actuators in unhealthy status exposure when the observing parameters exceed the threshold. Beeper and vibration motor are considered as the warner means in this prototype (Figure 2). The driver is located at the bottom of prototype and both beeper and vibrating motor are connected to the board through wire. The primary idea is to target the office working of the users who are more often are working in quiet places with the beeper. The beeper is positioned in the surface of the wearable toward the user for better sound quality. The beeper is quite tangent with the device housing box. In this design in fact, micro vibration as a complementary actuator to get the attention of group of people who are working in noisy environment and my not draw the attention in risky situations. The vibrating motor is located at the bottom of the 3-d housing case and touches the bone of the forearm for haptic warning. The vibrating motor and beeper both are connected to the driver board and synchronized. The notification system is programmed in three states. In one state which all observing parameters are in the safe range, it is off and no signal is generated.

The working category of the system is two modes while depending on the degree of risk. While the threshold is exceeded but not in extremely dangerous condition, the warning signal for actuators activation is created in low frequency. This frequency

and duration of signal activation is extended in highly risky situation to attract the user attention.

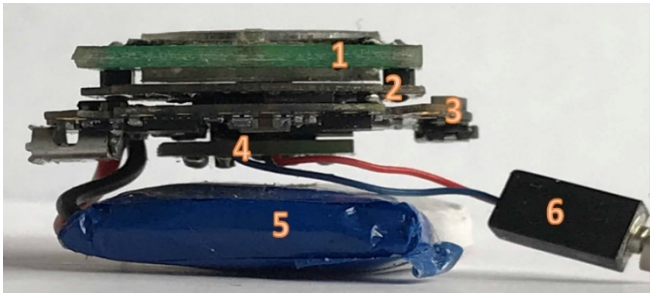


Figure 2 Platform, notification system layer, integrated and add on-board sensor of MLMS-EMG-3.2. (1)(2): gas sensor node consist of (1) sensor, and (2) gas sensor driver. (3): host platform, (4): notification driver, (5): battery, (6): vibrating motor connected to the driver.

2.2. Sensor Node's Specifications and Features

In this subsection, the most important features of the wrist-worn that provides some specific capabilities are bolded. This version of the prototype is implemented in three physical layers including one layer as the gas sensor node, host platform consists of integrated sensors and the bottom layer dedicated to the notification system driver (Figure 3, Figure 4). The MLMS-EMG-3.2 is provided as a wrist-worn in small size (42×42×25mm) and quite light (51.18 gr.). This box includes the platform, two layers at top and bottom and a LiPo / Li-Ion battery (Figure 4) in addition to beeper and vibrating motor that are located in the case's wall. Powering on/off of MLMS-EMG-3.2 is possible through a power accessible pin at the right side. In addition, the LiPo / Li-Ion battery is directly rechargeable with a USB window that is placed at the left side of the device. These specifications are limited to the physical design. The software implementation is discussed in different subsection.

- Ubiquitous device for hazardous gas detection: the flexibility, wearability, ease of use and cost of this wrist-worn, make it suitable to be applied by anyone in anywhere to detect anything (here hazardous gas). Usually, the type of hazardous gases in a specific environment are known to the user. (e.g. chemical laboratories). Many gas sensors operation at the same time (while might not be necessary), increase the power consumption and expand the size of the device by deploying number of gas sensors. One solution to reduce the difficulties to the least is to dedicate the sensor selection to the user. If the wearer can decide for the target sensor at any time and place, the efficiency of the design is improved.

gains. Whatever, the issues are pushed to the software side to be handled, the bottleneck dealing is more flexible. Consequently, the configuration is implemented in software side, filter and gain adjustment are performed manually on the hardware side. To the best of our knowledge, the best candidate which provides a complete solution is a Front End Amplifier (AFE) LMP91000 with very low power consumption. The LMP91000 is a micro-power AFE operating in the range of 2.7 to 5.25 V (in the MLMS-EMG-3.2 is used in 3.3v). The total current consumption with appropriate programming through I2C could be less than 10 µA [19]. LMP91000 amplifier converts the output current of the sensor to voltage proportionally and feeds to ADC of the microcontroller. LMP91000 is configurable for each target gas via I2C communication. In addition, a board to board connector in milli-scale also is soldered at the bottom of the sensor node. Bias $V_{ref.}$, R_{TIA}, RE are configured and unique for each gas sensor (Figure 2). Another factor for efficient performance is filter design. Depending on the sensor the filter design is varied. In general, a low pass and an integrator filter are implemented.

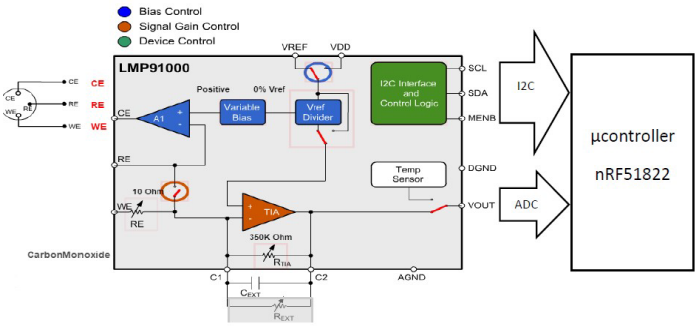


Figure 3 The internal architecture of LMP9100 and the communication with the host platform.

- Gas sensor node replaceability: to enable the sensor selection by the user, gas sensor node "replaceability" is considered as the solution. The gas sensor node that is located at the top of device through board to board connector, can be replaced by a different target gas sensor node easily. This can dramatically extend the number of gases that may be monitored (one at each time). By this feature, the device flexibility in terms of number of detected gases and wearability are improved and the power consumption is reduced. However, each gas sensor has its own configuration and requirements. The gain, filter and initialization of each sensor is customized. The number of utilized enable pins in the microcontroller are restricted due to the 24-pin board to board connector. As a conclusion, the hardware limitation is transferred from hardware side to software side for the more flexible solutions.

2.3. Device configuration and Sensor Selection

In the previous system version, the switching between the gas sensors was realized manually by the wearer. Each gas sensor was

dedicated an enable pin; the enable pin was point to the gas sensor configuration and initialization. This was demonstrated in hardware. Due to limited number of pins and future integration of other sensors (other than gas sensors), the pin utilization must be carefully performed. MLMS-EMG-3.1 version was in the scope of a one-way communication (data transmission) with a smartphone. In this version this communication is improved to bidirectional, where data are sent to the smartphone and also commands are received from the smartphone sent by the user.

Each command is sent through pressing the button on the smartphone screen. These buttons are implemented on the same screen as data observation, at the bottom.



Figure 4 Top and bottom side of MLMS-EMG-3.2 and data transition to android application.

2.4. Data packet transmission to the smartphone

In MLMS-EMG-3.2 data are categorized and transmitted in 3 separate packets (Figure 5): The data that are correlated due to transient efficiency and power consumption are combined and sent in one packet. For each packet different ID's have been assigned for identification.

- F1: motion elements, 19 bytes (including 2bytes don't care): For more accuracy, this packet is sent through BLE every 20ms. A 3DoF accelerometer, 3DoF gyroscope and 3DoF magnetometer are integrated for a precision user tracking. Each sensor reserves 12 bits (1.5 byte) which are highlighted with different colors. The packet F1 is started with Seq. which is used in data number counting as well as data logging. F1 is a data ID causing identification, byte separation and recognition process in android APK. The last two bytes are reserved for further development (see Figure 6).
- F2: environmental parameters (temperature, air pressure, and humidity), 12 bytes. This packet (see Figure 13b) is sent to the smartphone every second. In the same policy, it is started with seq. and followed by F2 and then air pressure, temperature and humidity are the parameters which are measured (Temperature also is possible to be measured through integrated sensor in humidity and pressure sensor through adequate configurations). All environmental elements are using 2 bytes divided to two higher and lower byte. Temperature data is represented also in signed 16-bit format, and the corresponding Celsius value can be computed by dividing the value with 256. Air humidity is also 16-bit value, but it is unsigned and the resulting value shows the relative humidity in percent (RH %). Air pressure uses also 16-bit representation, but the value is unsigned and the value must be divided by 32 to be converted into m bars.
- F5: Sound level, 4 bytes: The sounds packet construction is very similar to gas (see Figure 13d). However, the sound level detection is discussed in this paper and is a part of future development of the MLMS-EMG-3.2.

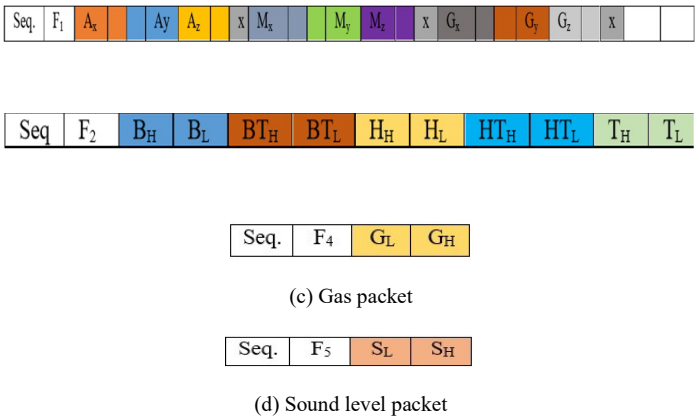


Figure 5 Packet structure of data transmission to the smartphone.

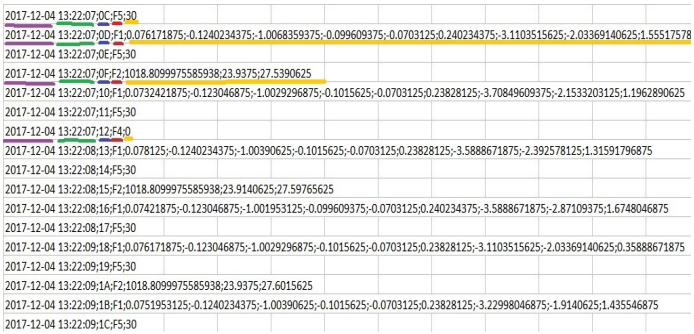


Figure 6 Date, time stamp, sequence, ID and value of each packet

2.5. Filter design

One of the most critical and challenging parts of this work is filter design. The filter must be implemented to eliminate the

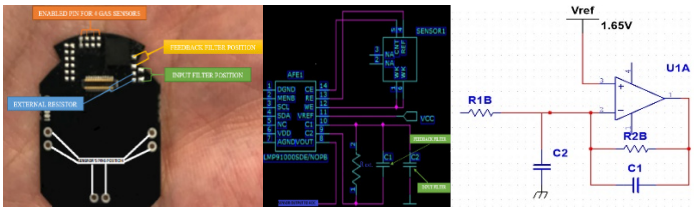
unpredicted and undesired signal noises from different sources as well as pass the real signals from the gas sensor. The filter in this project is a two-step Low Pass Filter (LPF-passive and active both are integrated in the circuit) as well as code implementation which forms a strict LPF with adequate cut-off frequency (Figure 8).

The first filter (input filter) includes a loaded resistor seen by the sensor and a capacitor (first filter including R_{1B} and C_2-passive). This resistor plays a crucial role in this circuit. It must be adjusted between 10-100 Ohm in the way that from one side it accomplishes the low pass filter design and on the other hand holds the gain in an acceptable range. R_1 is a tradeoff between response time and noise. Indeed, any noise on the WE probe is amplified by the op-amp with a gain of:

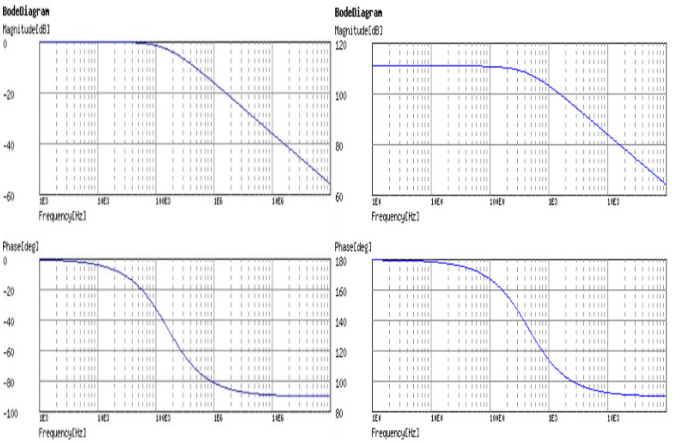
$$A = 1 + (R_{TIA\ (R2B)}/R_{1B})) \quad (1)$$

The second filter step is carried out through a strict active LPF. The real data from the sensor passed the first LPF are fed to the amplifier which amplifies the current signal as well as filters the noises with high frequency. This filter is a feedback filter which cut-off uniquely is function of C_1 and R_{2B}. This C_1 capacitor improves the stability too. Figure 9 shows the results from first LPF and second feedback filter. The combination of both filters is simulated in Multisim simulator and results are shown in the Figure 10.

The combination of these components is working well and has been simulated in a wide range of frequency (1 Hz – 270 kHz). In higher frequencies, almost all noise signals are eliminated properly, the phase and amplitude change of the output signal vs frequency is shown in Figure 10. Each gas sensor depends on the environment parameters, sensor calibration and sensor application must be uniquely studied. It is remarkable that, first step filtering does not have to be very narrow pass band than second one. PCB design of the sensor driving board and schematic are shown in Figure 7.



2.6. Calibration



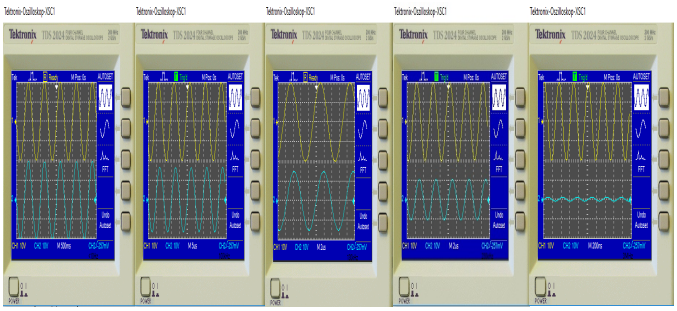


Figure 9 Results of filter application on gas sensor.

```
The target gas sensor is physically located;
The command for sensor configuration is sent to the device by the user;
The command is handled by the MLMS-EMG-3.2 and the related gas sensor node
configuration is initialized;
            VB is set;
             VB=1.65;
   Time interval of gas sensor output reading is set to 5 s;
          V output_is_read;
        While (VB-Voutput=>15 mv)
  {
     Read V output ();
        }
          Calibration process is started;
           Read temperature ();
          Read humidity ();
        Read pressure ();
Calibration is applied to the raw data;
          if (10mv <VB-V output< 15mv)
        {
          The status is detected as abnormal with degree of risk 2;
             The driver of notification system is activated;
             else if (5 mv <= VB-V output <= 10 mv)
         {
         The status is detected as abnormal with degree of risk 1;
              The driver of notification system is activated ();
          }
            The raw data are converted to ppm ();
```

Figure 10 MLMS-EMG-3.2 gas operation algorithm.

particular gas sensor, configuration and environmental parameters no more than 15 mV changes ($V_B - V_{output}$) can be achieved through gas concentration reaction with sensor membrane (otherwise it is not considered as a real signal but noise). If the V_{output} changes > 15 mV, this sample is rejected and immediately another sample is taken (Figure 11). If not, it is going through a calibration process (it's described in a separated paper). This procedure is followed by the second question on V_{output} change due to the degree of risk and two modes of notification system. As can be seen in the algorithm, the notification system configuration is strictly a function of gas concentration. At the end, raw data are

converted to ppm values. The final value of the target gas might be measured at the end of sensor stability (meanwhile in each sampling it is possible that the data reading is getting closer to the real value), but environmental evaluation status is notified to the user meanwhile (see Figure 10).

Carbon monoxide (CO) is a colorless, odorless, and tasteless toxic gas that is slightly lighter than air. The toxicity of CO is variable according to the concentration (Table 1). Long term exposition causes serious health issues. Therefore, measuring the CO concentration with a wearable device equipped with a notification system leads to an early individual notification and exposure avoidance. The focus of this work currently is on CO and NO_2 detection. The CO sensor is replaceable with NO_2 gas sensor.

Table 1 Risk of carbon monoxide in different concentrations [18].

Concentration(PPM)	Effect
0.1	Natural atmosphere level or clean air
50+	For a 50 ppm and higher CO is toxic for adults
70-75	Heart patients experience an increase in chest pain in this range. (HbCO 10%).
100	Headache, tiredness, dizziness, nausea is of signs of 100 ppm CO within 2 hrs of exposure. (Lewey & Drabkin)
200	Headache, nausea occur at this level. (NIOSH & OSHA)
400	life threatening within 3 hours
800	Healthy adults will have nausea, dizziness and convulsions within 45 minutes. Unconscious within 2 hours then death (determined in 1930)
1600	Headache, tachycardia, dizziness and nausea within 20 minutes

3.1. Experiment conditions

The experimental results presented in this paper are demonstrated under the controlled conditions in the chemical laboratory in center for life science automation (celisca) (air humidity: 59.1%, temperature: 22.7 °C, and air pressure: 1020.3 mbar). All results are produced from the prototype descripted above in the same laboratory and under the same conditions. For calibration, reliability and accuracy test three different gas sensor node layer were applied with different calibration factor (CF). MLMS-EMG-3.2 and BW clip as the calibrator were frequently exposed to the target gas and depending on the gas, the results were recorded in the specific time interval. The test container was a small closed transparent one. These experiments were experienced for 200 times a regular time interval within a week.

3.2. Device configuration

CO and NO_2 are two toxic gases from different families. CO is a reducing gas. When the gas sensor is exposed to CO, the internal resistor is reduced and consequently the output voltage is increased proportionally to the gas concentration. From these statement, it is understandable that, to detect the maximum range of gas concentration and avoid saturation (wider span), the LMP91000 of the sensor driver should be configured and set to the minimum initial voltage value. The deployed gas sensor is overloaded at 5,000 ppm. To cover this relatively large range, the LMP91000 is configured to:

$$V_B = \frac{1}{5} V_{ref} \quad (2)$$

$$\text{Where: } V_{ref} = 2.8\ v$$

$V_B = 0.56\ v$, indicates a gas concentration of 0 ppm. With $\mathbf{V_{ref}}$ (external source voltage) and $\mathbf{V_B}$, the difference $V_{ref} - V_B = 2.24v$ should cover the range of 0 to 5,000 ppm in CO concentration. According to the Table I, the CO measurement at very high concentrations (>1,700 ppm) does not help in individual rescue due to instant danger. In this paper, the experiments are limited to 1,700 ppm (majority of CO detectors are limited to 300 ppm). Consequently, R_{TIA} and also variable bias are configured. In Figure 2, the current flowing through R_{TIA} is called I_{sense}. Therefore, the V_{output} is a function of:

$$V_{output} = V_B\ (0\ ppm) + R_{TIA}.I_{sense} \quad (3)$$

The ADC is reading the V_{output} generated by the gas sensor. This value is in the range of:

$$V_B < V_{output} < V_{ref} \quad (4)$$

Generally, in design, calculation and calibration of the gas driver sensor the polarity of I_{sense} is a significant factor in (2). For reducing gases including CO, this polarity is always positive in the presence of target gas, thus V_B is expected to increase with the gas concentration. The opposite scenario is the case for oxidizing gases (e.g. NO_2). The sign of I_{sense} is negative and the respected generated voltage is reduced from the V_B. Consequently, V_{output} is proportionally smaller at higher concentrations. MLMS-EMG-3.2 has been exposed to different gas concentrations from 0 to 1,620 ppm. Within this range, performance, response time, recovery time and accuracy of the device are evaluated. The resolution of CO is categorized into two sections: 7.5 ppm (for CO concentrations < 50 ppm) and 5 ppm (for CO concentrations up to 1,620 ppm). All configurations, setting values for two groups of gas sensors, calibration process and conversion of each gas sensors are embedded in source code. Initialization for each target gas sensor is realized by sending the related command from the smartphone. During the experiments, several significant observations were perceived which are summarized as the follows:

- At high concentrations, when the different volume of gas is released into the container, the experiment for the gas with higher concentration is reaching to the some specific value in a faster rate rather than a test with lower concentration and volume of gas (both sensors reach to this point at the different time slopes). To differentiate between experiments, the time interval for the recording of the concentration must be at the same time interval, but before saturation state. In observation sections figures, number of portions are seen in each figure with the slop rate of response time or recovery time within the test (Figure 6, Figure 11, and Figure 13).
- During the measurement two status were observed in data recording: stable state and temporary unstable state. Stable state is a safe level that the recorded values are constant and reliable. In spite of stable state, the second state is consist of

gas values which are quickly recorded and disappeared. In fact, in this status a swinging values of measurement is recorded. Approximately every 50 ppm, the stable state was observed. This is experienced more often in higher concentration due to high volume of released target gas. In a normal exposure of target gas in low concentration, the response time might be as small as 15 second or less, while when the gas sensor is exposed to high concentration, the gas sensor does not have enough time for stability, therefore, the dynamic value updating in the form of swinging values are monitored. During the experiment the response time of 47 s was recorded for some tests, while this time might be more for the lower released gas volume. Halt time is increased with time and gas volume intensity and only occurs within stable points (Figure 12).

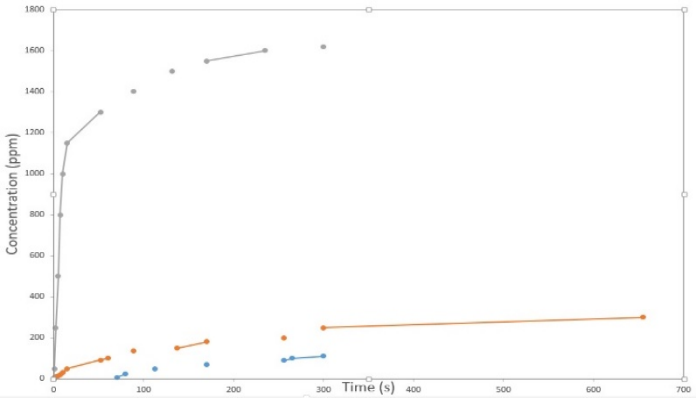


Figure 11 Response time for 110, 300 and 1620 ppm (CO).

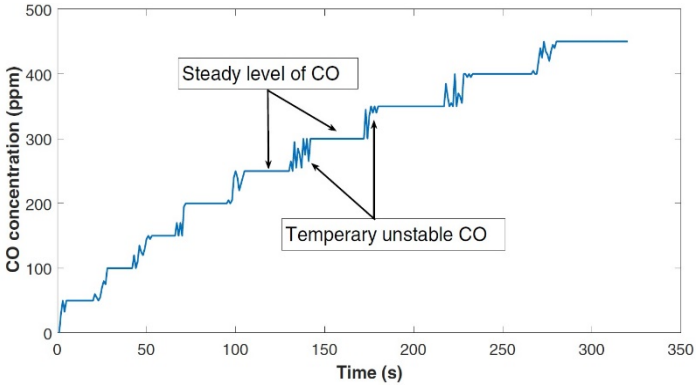


Figure 12 Stable level and halt time for 1,620 ppm (CO).

- The rate of response time is reduced gradually with the time (Figure. 8). The fastest response time is seen at the beginning of each experiment for the first measurement test. The rate of flowing gas into the container is constant for all experiments. To increase the concentration of the gas in experiments, the

- "Recovery time" for a gas sensor is defined as the essential required time for the sensor to reach from a non-zero ppm to zero ppm and is ready for the next experiment. The observation of several experiments in the laboratory, indicate after removing the sensor node from the gas exposure, for small period of time (seconds) the measured gas value is seen the same with the last change. Then a quick and large value drop is occurred. This drop, if followed with a sharper reduction of recorded values. The evaluation is continued to reach zero ppm but with slower rate than the previous stage. This pattern was repeated for three different gas concentrations (several experiments for each gas concentration), this is depicted in Figure 13.

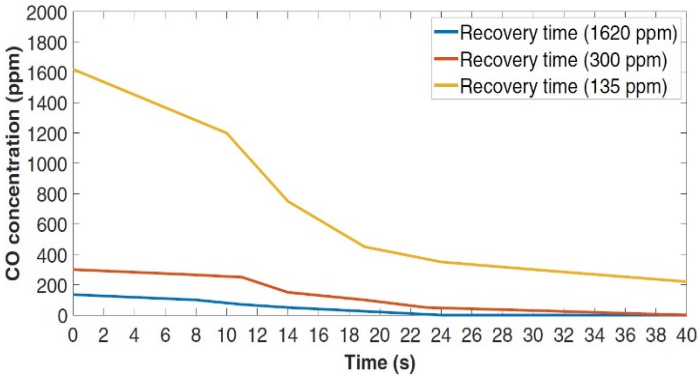


Figure 13 Recovery time for 135,300 and 1620 ppm (CO).

- Table 2, exhibits a limited number of recorded samples during the experiments under the different gas concentrations. These data are validated in conjunction well-known BW CLIP [21]. During the experiments MLMS-EMG-3.2 and BW CLIP were located into a container under the quite the same conditions. The maximum recorded error in the experiments is 6.5 %. Care must be taken that, the accuracy of the measurements are strongly the function of the "when the data are read". The authors would recommend for the average time of five minutes for the data assessment. The mean time of each test in our experiences were five minutes, however this time in some individual tests could be extended up to ten minutes. In the majority of the cases, the recorded gas concentration reaches to 75-80 % of the final value at the first five minutes.

Table 2 Experiment results of calibrator and MLMS-EMG-3.2.

MLMS-EMG-3.2 (ppm)	45	67.5	105	115	150	302
Calibrator (ppm)	42	65	111	126	152	300

4. Conclusion

In this work a compact, small and light weighted wrist-worn device for hazardous gas measurement was introduced. This prototype is based on physical layer design, named MLMS approach. Host platform including the physical ambient air sensors, and motion tracking sensors, is the middle layer. The gas sensor node consisting of the gas sensor and sensor driver is located at the top and the notification driver which is activated in abnormal status is located at the bottom of the host platform. MLMS-EMG-3.2 is designed for a large range of applications, due to particular feature that the gas sensor node is replaceable with some other target gas sensors. The target gas sensor is configured through sending commands from the smartphone by the user. By using the commands from the smartphone to activate the target sensor, the separate enable pin for each gas sensor is eliminated. This leads to a reduction the complexity and restrictions in hardware; instead a more flexible solution is presented in software side. Future developments of this prototype will include noise (sound level) detection.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work is funded by the Ministry of Economics, Employment and Health of Mecklenburg Vorpommern, Germany under project number MV (TBI-V-1-127- VBW-044).

References

[1] Haghi, Mostafa, Kerstin Thurow, and Norbert Stoll. "A Multi-Tasking, Multi-Layer and Replaceable Wrist-Worn Environmental Monitoring Sensor Node." In 2018 5th International Conference on Control, Decision and Information Technologies (CoDIT), pp. 25-31. IEEE, 2018. 10.1109/CoDIT.2018.8394781

[2] E. of the European Parliament, of the Council of 21 May 2008 on ambient air quality, and cleaner air for Europe, 2008. http://ec.europa.eu/environment/archives/enlarg/handbook/air.pdf [Accessed on August 13, 2018].

[3] I. Kousuke, and J. Watson. *The Stannic Oxide Gas Sensor Principles and Applications*. CRC press, 2017. Town is missing

[4] D. Oletic, V. Jelicic, D. Antolovic, and V. Bilas, "Energy-efficient atmospheric co concentration sensing with on-demand operating mox gas sensor," in SENSORS, 2014 IEEE. IEEE, 2014, pp. 795–798. DOI: 10.1109/ICSENS.2014.6985119

[5] R. S. Istepanian, E. Jovanov, and Y. Zhang, "Guest editorial introduction to the special section on m-health: Beyond seamless mobility and global wireless health-care connectivity," IEEE Transactions on information technology in biomedicine, vol. 8, no. 4, pp. 405–414, 2004. DOI: 10.1109/ACCESS.2015.2437951

[6] M. Haghi, K. Thurow, and N. Stoll, "A multi-layer multi-sensor wearable device for physical and chemical environmental parameters monitoring (co & no_2)" in Information and Digital Technologies (IDT), 2017 International Conference on. IEEE, 2017, pp. 137–141. DOI: 10.1109/DT.2017.8024285

[7] M. Haghi, K. Thurow, and R. Stoll, "Wearable devices in medical internet of things: scientific research and commercially available devices," Healthcare informatics research, vol. 23, no. 1, pp. 4–15, 2017

[8] I. Stassen, N. Burtch, A. Talin, P. Falcaro, M. Allendorf, and R. Ameloot, "An updated roadmap for the integration of metal–organic frameworks with electronic devices and chemical sensors," Chemical Society Reviews, vol. 46, no. 11, pp. 3185–3241, 2017. DOI: 10.1039/C7CS00122C

[9] D. Oletic and V. Bilas, "Design of sensor node for air quality crowd sensing," in Sensors Applications Symposium (SAS), 2015 IEEE. IEEE, 2015, pp. 1–5. DOI: 10.1109/SAS.2015.7133628

[10] R. R. Fletcher, N. M. Oreskovic, and A. I. Robinson, "Design and clinical feasibility of personal wearable monitor for measurement of activity and environmental exposure," in Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE. IEEE, 2014, pp. 874–877. DOI: 10.1109/EMBC.2014.6943730

[11] F. Sanfilippo and K. Y. Pettersen, "A sensor fusion wearable health monitoring system with haptic feedback," in Innovations in Information Technology (IIT), 2015 11th International Conference on. IEEE, 2015, pp. 262–266. http://doi.ieeecomputersociety.org/10.1109/INNOVATIONS.2015.7381551

[12] M. Haghi, K. Thurow, and N. Stoll, "A three-layer multi-sensor wearable device for physical environmental parameters and NO_2 monitoring," in Smart Systems and Technologies (SST),2017 International Conference on. IEEE, 2017, pp. 149–154. DOI: 10.1109/SST.2017.8188686

[13] M, Haghi, R. Stoll and K. Thurow, 2018. A Low-Cost, Standalone, and Multi-Tasking Watch for Personalized Environmental Monitoring. *IEEE transactions on biomedical circuits and systems*, (99), pp.1-11. 10.1109/TBCAS.2018.2840347

[14] http://site.jjstech.com/pdf/GAExtremeManual.pdf. [Accessed on August 13, 2018].

[15] http://www.ti.com/lit/ug/tidub41/tidub41.pdf. [Accessed on August 13, 2018].

[16] http://www.ti.com/lit/ug/snoa922/snoa922.pdf. [Accessed on August 13, 2018].

[17] H. Cho, "Personal environmental monitoring system and network platform." *Sensing Technology (ICST), 2015 9th International Conference on*. IEEE, 2015. DOI: 10.1109/ICSensT.2015.7438496

[18] https://iprotoxi.fi/ [Accessed on 10October 02, 2018].

[19] http://www.ti.com/lit/ds/symlink/lmp91000.pdf. [Accessed on August 13, 2018].

[20] https://www.spec-sensors.com/wp-content/uploads/2016/05/Carbon-Monoxide-Risks-at-Low-Levels.pd. [Accessed on August 13, 2018].

[21] http://www.honeywellanalytics.com/en/products/BW-Clip. [Accessed on August 13, 2018].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 252-256 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Frameworks for Performing on Cloud Automated Software Testing Using Swarm Intelligence Algorithm: Brief Survey

Mohammad Hossain*, Sameer Abufardeh, Sumeet Kumar

Department of Math Science and Technology, University of Minnesota Crookston, Crookston, MN 56716, USA.

ARTICLE INFO

Article history:
Received: 01 March, 2018
Accepted: 16 March, 2018
Online: 11 April, 2018

Keywords:
Swarm Intelligence
Ant Colony Algorithm
Bee Colony Optimization
Cloud Based Testing
Test Case Prioritization

ABSTRACT

This paper surveys on Cloud Based Automated Testing Software that is able to perform Black-box testing, White-box testing, as well as Unit and Integration Testing as a whole. In this paper, we discuss few of the available automated software testing frameworks on the cloud. These frameworks are found to be more efficient and cost effective because they execute test suites over a distributed cloud infrastructure. One of the framework effectiveness was attributed to having a module that accepts manual test cases from users and it prioritize them accordingly. Software testing, in general, accounts for as much as 50% of the total efforts of the software development project. To lessen the efforts, one the frameworks discussed in this paper used swarm intelligence algorithms. It uses the Ant Colony Algorithm for complete path coverage to minimize time and the Bee Colony Optimization (BCO) for regression testing to ensure backward compatibility.

1. Introduction

With the increase of code complexity in modern software, chances of errors are exponentially increasing. These errors can cause loss of money and innocent human lives [1]. For example, in April 24 1994, a China airline airbus A-300 crashed, due to a software bug, resulting the death of 264 innocent lives [2, 3]. Another famous incident was reported in April 1999, where a small software bug in a military satellite was behind a $1.2 billion loss: one of the costliest unmanned accidents in the history of Cape Canaveral launches [4, 5].

Therefore, to reduce the risks of errors, researchers have developed a variety of testing techniques to find and fix software bug early and before the deployment of the software. One of the most critical tests is unit testing, where each module of program is tested separately. Another critical test is integration testing that occurs after unit testing, where individual software modules are combined and tested as a group. Since unit testing requires access to the system code, it is done during the initial stages of a program, detecting an estimated 65% of the errors [6, 7, 8]. Other types of tests includes, system testing and acceptance testing. In system testing, the system is tested as a whole to verify that it meets the specified requirements. After that, acceptance testing is done to verify that, the system meets the client/user requirements.

*Corresponding Author: Mohammad Hossain, Crookston, MN 56716, 218-281-8222, Email: hossain@crk.umn.edu

Testing is generally a lengthy and costly process. Therefore, automated software testing generally intended to reduce the time and the cost of testing. It also can increase the depth and scope of tests to help improve software quality. However, most automated testing tools fails to provide efficient results, because they only focus on specific testing techniques [9, 10] and in sometimes they may be unsuitable for large-scale software.

This survey paper focuses on testing based on Cloud platforms tools in order to develop cost effective, efficient and time saving tools that follows the rules of research based techniques to produce software free of or with few errors or bugs.

2. Testing

Software testing is an investigation conducted to provide stakeholders with information about the quality of the product or service under test. Software testing can also provide an objective, independent view of the software to allow the business to identify and understand the risks of software implementation. Testing techniques include, but are not limited to, the process of executing a program or application with the intent of finding software bugs (errors or other defects). It is the process of validating and verifying a software program, application or product [11]. Software testing is a huge domain, but it can be broadly categorized into two areas: manual testing and automated testing:

2.1. Manual Testing

An individual or a group of individuals performing all of the software quality assurance testing, checking for errors and defects, is knows as manual testing.

2.2. Automated Testing

The main purpose of this testing is to replace manual testing with automated cloud based testing without loss of efficiency, in such a way that it can not only save time but also produce high quality software.

3. Terminologies and Abbreviations

The following abbreviations and terminologies were used in this research paper.

3.1. Regression Testing

Software undergoes constant changes. Such changes are necessitated because of defects to be fixed, enhancements to be made to existing functionality, or new functionality to be added. Anytime such changes are made, it is important to ensure that, first changes or additions work as designed. Second changes or additions are something that is already working and should continue to work. Regression testing is carried out to ensure that any new feature introduced to the existing product does not adversely affect the current [11].

3.2. Traceability

Traceability is defined as the ability to describe and follow the life of a requirement, in both forward and backward direction, throughout the software life cycle. Traceability relations can assist with several activities of the software development process such as evolution of software systems, compliance verification of code, reuse of parts of the system, requirement validation, understanding of the rationale for certain design decisions, identification of common aspects of the system, and system change and impact [12,13].

3.3. White-Box Testing (WBT)

White-Box Testing, also known as clear box testing, gives verification engineers full access to the source code and the internal structure of the software. It is the detailed investigation of internal logic and structure of the code [14]. In WBT, it is necessary for a tester to have full knowledge of source code. Some important types of WBT techniques includes Statement Coverage (SC), where tester tests every single line of code, and Condition Coverage (CC) in which all the conditions of the code are checked by providing true and false values to the conditional statements in the code.

3.4. Black-Box Testing (BBT)

BBT treats the software as a “Black Box” without any knowledge of internal working of the system and it only examines the fundamental aspects of the system. In BBT the tester has knowledge of the system architecture but he/she does not have access to the source code [15].

4. Swarm Intelligence Algorithms to Optimize Regression Testing.

Regression Testing ensures that any changes or enhancement made to the system will not adversely affect the functionality of software. The execution of all test cases can be an costly and time consuming process. With this in hand, prioritization of test cases can help in reduction in cost of regression testing. Swarm intelligence is an emerging area in the field of optimization and researchers have developed various algorithms by modeling the behaviors of different swarm of animals and insects such as ants, termites, bees, birds, fishes [16]. These algorithms are being used to reduce the time and cost of testing in general and more specifically regression testing [11, 17]. Two such widely used algorithms are Ant Colony Algorithm and Bee Colony Optimization.

4.1. Ant Colony Algorithm (ACA)

Ant colony algorithms are based on the behavior of a colony of ants when looking for food. In their search they mark the trails they are using by laying a substance called pheromone. The amount of pheromone in a path tells other ants if it is a promising path or not. This observation inspired Colorni, Dorigo and Maniezo [18] for proposing a metaheuristics technique: ants are procedures that build solutions to an optimization problem. Based on how the solution space is being explored, some values are recorded in a similar way as pheromone acts, and objective values of solutions are associated with food sources. An important aspect of this algorithm is parallelism: several solutions are built at the same time and they interchange information during the procedure and use information of previous iterations [19]. In [20] Li and Lam proposed to use UML State chart diagrams and ACA for test data generation. The advantages of their proposed approach are that this approach directly uses the standard UML artifacts created in software design processes and it also automatically generated feasible test sequence, non-redundant, and it achieves the stated coverage criteria.

4.2. Bee Colony Optimization (BCO)

Bee swarm behavior in nature is characterized by autonomy and distributed functioning, and it is self-organizing. Recently, researchers started studying the behavior of social insects in an attempt to use the Swarm Intelligence concept in order to develop various Artificial Systems [21]. In software engineering, BCO can be used as a method of regression testing and traceability. Its purpose is to verify that the current version of the software is compatible with pervious test results and the data is comparable to previous versions of the software.

Karnavel and Santhoshkumar proposed a fault coverage regression system exploiting the BCO algorithm discussed in [11]. The idea is based on the natural bee colony with two types of worker bees that are responsible for the development and maintenance of the colony: scout bees and forager bee. The BCO algorithm developed for the fault coverage regression test suite is based on the behavior of these two bees. The algorithm has been formulated for fault coverage to attain maximum fault coverage

in minimal units of execution time of each test case. Two examples were used whose results are comparable to the optimal solution [22, 23]. The system was divided into the following components: *Testing Phase, Traceability Phase, Exploration Test, Generating Reports, and Storing in Database* (figure 1):

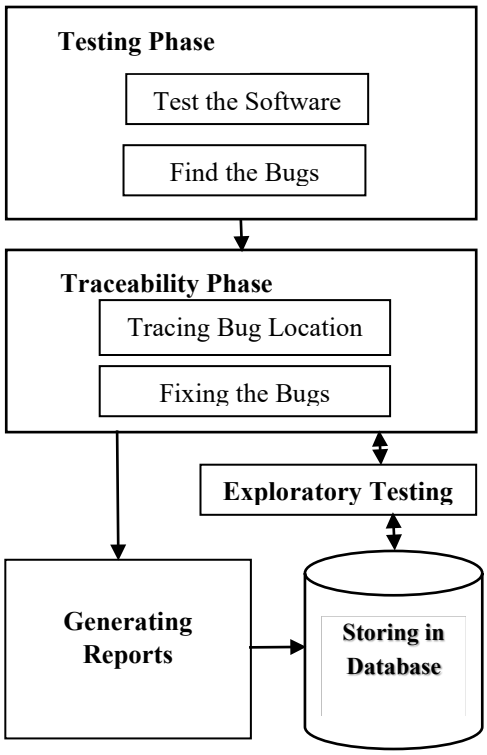


Figure 1: Architecture of BCO based fault coverage regression test.

5. Cloud Testing Framework Example

Large software requires a huge number of test cases. Often, these test cases require a lot of time and effort even with automated testing [24, 25, 26]. Because each test consumes execution time, the execution time required generally decreases when parallelization is used. Cloud Testing Frameworks presented in [24, 26] improves the execution process time by performing the parallel execution of test cases using current computational resources without source code modification [24, 26, 27].

Using a distributed and parallelized test can result in significant reduction of time required for test cases execution. It can also reduce the needed to identify and correct faults. Hence, reducing the total cost of development. Furthermore, the proposed framework in [24] increases the reliability of the test results by using heterogeneous environment that can result in the exposure of hidden failures ahead of the production phase [24].

One of the earliest Cloud Testing Framework "*CouldTesting*" presented by Oliveira and Duarte in 2013 is shown in figure 2. The framework distributes the unit tests using reflection on the local classes and then schedules the machine on cloud. It then the load is distributed over machines, using the round robin scheduling algorithm to ensure even distribution of the requests to the available machines in the test infrastructure [24].

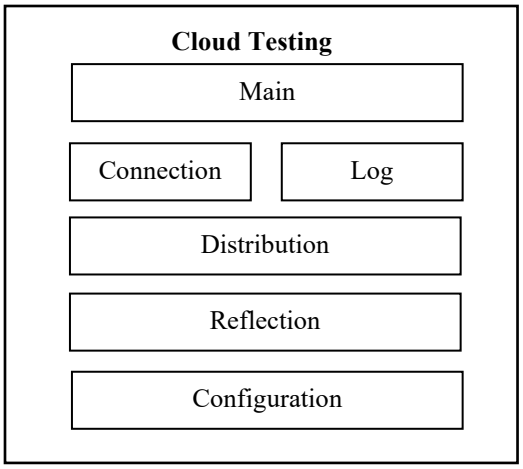


Figure 2: Cloud testing components from [24]

Figure 2 presents the architectural components of the framework. The main components of the framework are: Configuration, Reflection, Distribution, Connection, Log and Main.

The *configuration* component help in defining information for *paths, hosts,* and for *load balancing* [24]. This component deals with issues related to local storage space allocation for test results, selecting the libraries needed for proper test execution, and file access permission. It includes the list of machines and the parameters for the load balancer. The Reflection component extracts the tests cases [24].

Figure 3 depicts the distribution component being used to intermediate the execution of test suites over a parallel infrastructure. To work with a given IDE and parallel infrastructure the framework must be extended to include specific plugins. The *connection* component provides an interface on the client side to communicate with the cloud provider. In the cloud site this component provides a service that manages the execution of each test and it sends the test results back to the client in real-time. The log component records events generated in the process. The main component is a facade that encapsulates the components [24].

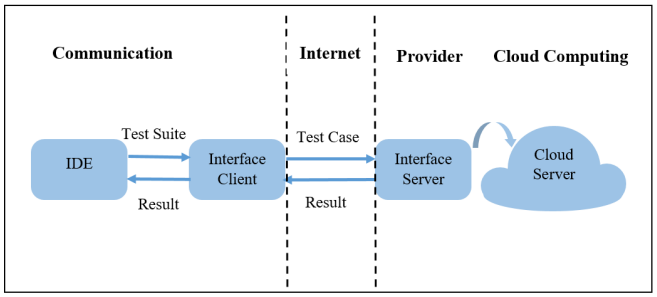


Figure 3: Distribution component being used to intermediate the execution of test suites over a parallel infrastructure from [24].

6. Test Case Prioritization Techniques

With the limited testing budget, it has been always a big challenge of software testing to optimize the order of test case execution in a test suite, so that they detect maximum number of errors. Three solutions to this problem were discussed in [28] such as test suite reduction, test case selection, and test case prioritization.

As the name suggests, the test suite reduction removes test cases from a test suite, which are redundant, and test case selection selects the most fault revealing tests based on a given heuristic. On the other hand, Test Case Prioritization (TCP) considers all the test cases without removing any of them. Instead it ranks all existing test cases thus prioritizes the test cases. While testing is performed testers executes the test cases with the higher priorities first as long as the testing budged supports [25]. Following techniques are discussed in [25, 28] to implement TCP.

6.1. Code/Topic Coverage Based

Most popular technique that has been reported in the literature is the Code coverage-based TCP that prioritize the test cases effectively [29]. It requires knowing the source code information of the software to measure the code coverage. This technique is not applicable in black- box systems tests because of lack of information about the code coverage. In [30] Thomas et. al proposed a modified a TCP technique, which they called topic coverage, provides an alternative concept to code coverage. The goal was to rank tests so that they cover more topics sooner [25, 28].

6.2. Text Diversity Based

Another common technique for TCP is to diversify the test cases. In [25] authors described a test case diversifying techniques that analyzes the test scripts directly hence this approach is called a text diversity-based TCP. The technique treats test cases as single, continuous strings of words. Then it applies different string distance metrics, such as the Hamming distance, on pairs of test cases and determines their dissimilarity. The idea is that the more two test cases are dissimilar textually the more they are likely to detect faults in different part of the source code [25, 28].

6.3. Risk Based Clustering

This TCP needs to access to execution results of the previous test cases, typically examining only the last execution of the test cases. It can be extended to as many previous executions as possible. This technique must ensure to run those test cases that failed in their previous execution provided that they are still relevant [31]. The technique might be combined with other TCP techniques, as well. For example, one can prioritize the previously failed test cases using a coverage-based approach to provide a full ordering of the test cases. In [28] authors modified this approach to have several clusters of test cases of different riskiness factor rather than having only two clusters of failed and non-failed test cases. In their approach, the highest risk is assigned to the tests that failed in the immediate version before the current version. The next riskiest cluster are tests that did not fail in the previous version but failed in the two versions before the current version, and so on.

7. Proposed Frameworks Discussed

As we discussed earlier, Swarm Intelligence algorithms as an emerging technique in the field of optimization are being integrated in many of the automation testing frameworks. These algorithms were instrumental in improving the performance and the efficiency of software testing on the cloud. In this section, we briefly highlight few of the frameworks that used Swarm Intelligence algorithms:

In [23] A. Kaur and S. Goyal proposed a Bee Colony Optimization algorithm for fault coverage-based regression test suite prioritization. The framework imitates the behavior of two types of worker bees found in nature. The behavior of the bees has been observed and mapped to prioritize software test suite.

A similar system has been presented by Karnavel and Santhoshkumar in [11] that used Bee Colony Optimization algorithm for test suite prioritization. The authors modified an existing framework to reduce the number of test cases from the retest test pool.

G. Oliveira and A. Duarte presented one of the aerialist frameworks called 'CloudTesting'. The framework executed test cases in parallel over a distributed cloud infrastructure [24]. In the framework, cloud infrastructure is used as the runtime environment for automated software testing. Their experimental results indicate remarkable performance gain without significantly increasing the cost involved in facilitating the cloud infrastructure. The framework also simplified the execution of automatic tests in distributed system.

The framework presented by S. Faeghe and S. Emadi in [32] used the Ant Colony algorithm to automate software path test generation. The authors proposed a solution based on ant colony optimization algorithm and model-based testing for faster generation of test paths with maximum coverage and minimum time and cost. According to authors' evaluation, the framework showed better performance over the existing methods in terms of cost, coverage and time.

A. Kaur and D. Bhatt in [29] proposed a regression testing based on Hybrid Particle Swarm Optimization (HPSO). The HPSO is an algorithm where a Genetic Algorithm (GA) is being introduced to the Particle Swarm Optimization (PSO) concept. The authors used the hybrid approach to prioritize tests for regression testing. The authors also mentioned that use of algorithm improved the effectiveness of their proposed framework.

8. Conclusion

Testing is one of the most complex and time-consuming activities. Automated testing on the cloud is one of the most popular solutions to reduce the time and the cost of software testing. In this paper, we discussed examples of automated frameworks proposed for testing software on the cloud [11, 23, 24, 29, 32]. Based on our review to such frameworks, it is evident that the use of the cloud as runtime environment for software testing are more efficient and effective solution when compared to traditional methods. Furthermore, on the cloud automated testing frameworks that used Swarm Intelligence Algorithms such ACA and BCO were able to produce significant reduction in the time required to execute large test sets and cover a diverse and heterogenic testing coverage. The use of such effective algorithms facilitates and enhances parallel execution and distribution of large testing loads on the cloud. In addition, cloud-testing frameworks generally simplifies the execution of automatic tests

in distributed environments, hence gains in performance, reliability and simplicity of configuration.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors would like to thank the reviewers of this work for their help. This work was accomplished by the support of University of Minnesota Crookston.

References

[1] R. Hower, "What are some recent major computer system failures caused by software bugs," http://www.softwareqatest. com/qatfaq1.html,[February 2015], 2005.
[2] P. Ladkin, "The crash of flight ci676." [Online]. Available:http://www.rvs.uni-bielefeld.de/publications/Reports/taipei/taipei.html
[3] P. Krömer, J. Platoš, V. Snášel, "Nature-inspired meta-heuristics on modern GPUs: state of the art and brief survey of selected algorithms." International Journal of Parallel Programming 42.5 (2014): 681-709.
[4] B. Beizer, Software testing techniques. Dreamtech Press, 2003.
[5] M. Mavrovouniotis, C. Li, S. Yang, "A survey of swarm intelligence for dynamic optimization: algorithms and applications", Swarm and Evolutionary Computation 33 (2017): 1-17.
[6] N. Gupta, "Different approaches to white box testing to find bug" International Journal of Advanced Research in Computer Sciency and Technology (IJARCST 2014) 2.3 (2014): 46-9.
[7] M. Nouman, U. Pervez, O. Hasan, K. Saghar, "Software testing: A survey and tutorial on white and black-box testing of c/c++ programs," in Region 10 Symposium (TENSYMP), 2016 IEEE. IEEE, 2016, pp. 225–230.
[8] B. Balamurugan, J. Sridhar, D. Dhamodaran, P. Venkata Krishna. "Bio-inspired algorithms for cloud computing: a review." International Journal of Innovative Computing and Applications 6, no. 3-4 (2015): 181-202.
[9] R. Khalid, "Towards an automated tool for software testing and analysis" In Applied Sciences and Technology (IBCAST), 2017 14th International Bhurban Conference on, pp. 461-465. IEEE, 2017.
[10] E. Pacini, C. Mateos, C.G. Garino. "Distributed job scheduling based on Swarm Intelligence: A survey." Computers & Electrical Engineering 40.1 (2014): 252-269.
[11] K. Karnavel1, J. Santhoshkumar, "Automated Software Testing for Application Maintenance by using Bee Colony Optimization algorithms (BCO)", Information Communication and Embedded Systems (ICICES), 2013 International Conference on. IEEE, 2013.
[12] U. Salima and A. Askarunisha, "Enhancing the Efficiency of Regression Testing Through Intelligent Agents" International Conference on Computational Intelligence and Multimedia Applications 2007, pp. 103-108.
[13] S. Dick, A. Kandel. Computational intelligence in software quality assurance. Vol. 63. World Scientific, 2005.
[14] M. E. Khan, F. Khan. "A comparative study of white box, black box and grey box testing techniques." Int. J. Adv. Comput. Sci. Appl 3.6 (2012).
[15] M. Khan, "Different Approaches to Black Box Testing Technique for Finding Errors," IJSEA, Vol. 2, No. 4, pp 31-40, October 2011
[16] K. Dervis, B. Akay. "A survey: algorithms simulating bee swarm intelligence." Artificial intelligence review 31.1-4 (2009): 61.
[17] A. Kaur, D. Bhatt "Hybrid Particle Swarm Optimization for Regression Testing", International Journal on Computer Science and Engineering, Vol. 3 No. 5 (May-11), pp1815~1824, ISSN: 0975-3397
[18] A. Colorni, M. Dorigo, V. Maniezzo. "Distributed Optimization by Ant Colonies". First European Conference on Artificial Life, 134-142, 1991.
[19] S. Mazzeo, and I. Loiseau. "An ant colony algorithm for the capacitated vehicle routing." Electronic Notes in Discrete Mathematics 18 (2004): 181-186.
[20] H. Li, P.L. Chiou, "Software Test Data Generation using Ant Colony Optimization." International conference on computational intelligence. 2004.
[21] D. Teodorovic, M. Dell'Orco. "Bee colony optimization–a cooperative learning approach to complex transportation problems." Advanced OR and AI methods in transportation (2005): 51-60.
[22] O. Gotel, and A. finkelstein "An analysis of the Requirements Traceability Problem", International conference on Requirements Engineering, USA, 1994
[23] A. Kaur and S. Goyal, "A Bee Colony Optimization Algorithm for Fault Coverage Based Regression Test Suite Prioritization", International Journal of Advanced Science and Technology Vol. 29, April, 2011
[24] G. Oliveira, A. Duarte "A Framework for Automated Software Testing on the Cloud", 2013 International Conference on Parallel and Distributed Computing, Applications and Technologies
[25] Y. Ledru, A. Petrenko, S. Boroday, and N. Mandran, "Prioritizing test cases with string distances," Automated Software Engineering, vol. 19, no. 1, pp. 65–95, 2011.
[26] A. Duarte, W. Cirne, F. Brasileiro andP. Machado, "Gridunit: software testing on the grid." In Proceedings of the 28th international conference on Software engineering (pp. 779-782). ACM, May 2006
[27] T. Banzai, H. Koizumi, R. Kanbayashi, T. Imada, T. Hanawa, M. Sato. "D-cloud: Design of a software testing environment for reliable distributed systems using cloud computing technology." In Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on, pp. 631-636. IEEE, 2010.
[28] H. Hemmat, Z. Fang, M. V. Mantyla , "Prioritizing Manual Test Cases in Traditional and Rapid Release Environments", Software Testing, Verification and Validation (ICST), 2015 IEEE 8th International Conference on. IEEE, 2015.
[29] G. Rothermel, R. H. Untch, C. Chu, and M. J. Harrold, "Prioritizing test cases for regression testing," Software Engineering, IEEE Transactions on, vol. 27, no. 10, pp. 929–948, 2001.
[30] S. Thomas, H. Hemmati, A. Hassan, and D. Blostein, "Static test case prioritization using topic models," Empirical Software Engineering, vol. 19, no. 1, pp. 182–212, 2014
[31] A. Onoma, W. Tsai, M. Poonawala, H. Suganuma, "Regression testing in an industrial environment," Communications of the ACM, vol. 41, no. 5, pp. 81–86, 1998.
[32] S. Faeghe, S. Emadi. "Automated generation of software testing path based on ant colony." In Technology, Communication and Knowledge (ICTCK), 2015 International Congress on, pp. 435-440. IEEE, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 254-259 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Wireless Channel Measurement and Modeling in Industrial Environments

Kun Zhang[1], Liu Liu[*,1,2], Cheng Tao[1], Ke Zhang[1], Ze Yuan[1], Jianhua Zhang[3]

[1]*School of Electronic and Information Engineering, Beijing Jiaotong University, 100044, P.R.China*

[2]*National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, P.R.China*

[3]*Beijing University of Posts and Telecommunications, 100876, P.R.China*

ARTICLE INFO

Article history:
Received: 08 June, 2018
Accepted: 03 August, 2018
Online: 05 August, 2018

Keywords:
IIoT
Wireless channel
K factor
Time delay
Path loss

ABSTRACT

The Industrial Internet of Things (IIoT), a typical use case of the Internet of Things, has been a great prospect for development in the near future. In the developing fifth generation mobile communications, the IIoT is an extremely important application case. In this paper, the channel propagation properties are researched and modeled based on the measurement data in an automobile welding factory. The results of the path loss exponent values are around 2.5 and 3.3 in line of sight propagation (LOS) and obstructed line of sight (OLOS) propagation scenarios, respectively. In LOS and OLOS propagation scenarios, K factor values are around 5 dB and 4 dB, respectively. Meanwhile, in these two-propagation environment, the amount of multipath components (MPCs) and the root mean squared (RMS) delay are extracted and compared. Particularly, a special case is considered that the sensor/actuator is placed inside the metallic body of a machine, which affects the path loss exponent values and K factor values. These measurement results are of great significance to the development of IIoT.

1. Introduction

In recent years, the Internet of things has become a hot research topic in industry and academia. In the developing fifth generation wireless mobile communications, the IIoT is an extremely important application case. The IIoT is particularly important in the application of the fifth generation of wireless communications. In the IIoT wireless communications, there are two different types of radio signals: narrow band (NB) and wide band (WB). In the industrial environment, there are a great quantity of mobile equipments work at WB to transmit surveillance video. Basically, there are an enormous amount of sensors work at NB to obtain all kinds of temperature, humidity or pressure information These signals also have different central frequencies. This paper mainly investigates the wireless channel of industrial environment and it is an extension of work originally presented in 2018 20th International Conference on Advanced Communication Technology [1].

The IIoT generally includes three different layers: application layer, perception layer and network layer. The first layer is to be used intelligent sensing technology, such as various sensors, to collect industrial data anytime and anywhere; the second layer is to be used communication technology, such as the Zig-bee, the Narrow Band Internet of Things (NB-IoT), and so on, to build a communication network. The third layers is to be used data processing and analysis technology, such as cloud computing, and so on, to make deep mining and utilization of these data, and then to realize the management and optimization of the industrial process [2]. At the IIoT network layer, huge amounts of sensors connect to the access point by wireless communication technologies. Wireless communication links will be obstructed and shadowed by industrial equipment or metallic materials in the industrial propagation environments. The propagation of electro-magnetic waves will generate strong reflection due to the role of these obstructions. However, in the industrial propagation case, wireless communication technology has not been completely researched. There are many differences between IIoT and cell communication systems. It is important to get a thorough knowledge of the IIoT wireless channel characteristics.

At present, researches on the influence of modern industrial environment on electromagnetic wave propagation are relatively

[*]Liu Liu, Beijing Jiaotong University No.3 Shangyuancun Haidian District Beijing China, Email: liuliu@bjtu.edu.cn

limited. Reference [3] extracted wireless channel parameters of path loss and time delay power based on the measured data of five factories. Reference [4] studied the characteristics of the typical industrial environment by channel measurement based on manufacturing plant. The related parameters obtained from the average time delay and RMS delay obtained from the 2.4GHz center frequency were found to be dependent on the LOS condition and related to the distribution of metal hindrance. Reference [5] measured four typical factory environments (corridors, offices, laboratories and large industrial manufacturing halls) at 1.9GHz center frequency, and analyzed shadow fading, K factor and delay spread which according with log-normal distribution. Reference [6] considered the large-scale characteristics based on measurement data in several factory buildings. Reference [7] researched NB propagation at 900 MHz, 2400 MHz and 5200 MHz. Reference [8] studied the effects of two important industrial environments on electromagnetic wave propagation, including channel path loss and multipath component (MPCs). One is high absorption environment, the other is high reflection environment. In the same year, reference [9] proposed a broad-band channel model for industrial wireless networks. This model takes into account the influence of noise in the harsh factory environment and uses the first order two state Markov process to describe the characteristics of the typical burst pulse noise in the industrial environment. Authors in reference [10] modelled the industrial wireless channels using Saleh-Valenzuela models. In 2017, reference [11] conducted seven scene measurements at three locations, studied the LOS and OLOS scenarios with a frequency of 5.8GHz, and analyzed the wireless channels.

The main contribution of this paper is to employ a vector signal generator and spectrum analyzer as a channel sounding system to probe realistic automobile factory environments. Three frequency bands, 1.1GHz, 2.55GHz, and 5.8GHz are considered in this paper. The frequency bands of 1.1GHz and 2.55GHz are corresponding to narrow band (NB) (0.8MHz) and 5.8GHz is corresponding to wide band (WB) (8MHz). In the NB case, we consider a special condition that the sensor/actuator (receiver antenna of the channel sounder in our channel measurements) is placed inside the metallic body of a machine, an industrial computer box with the thickness of 2 mm in the measurement. The channel parameters of time dispersion including RMS delay spread, number of MPCs, power difference and the K factor in different scenarios are extracted from the measurement data.

The paper is organized as follows. In Section 2, the system model and measurement system are described. Then raw data are analyzed to show the results of measurement and give the reasonable explanation in Section 3 . In Section 4 the conclusion is given.

2. System model and measurements description

As shown in Figure 1, a vector signal generator (Rhodes &Schwartz SMBV100A) and a spectrum analyzer (Agilent N9010B) are employed as a channel sounder. Two local rubidium clocks are employed to ensure efficiency and accuracy synchro-nization between the Tx and Rx. In Table 1 channel measurement parameters are listed.

As shown in Figure 3, two different propagation types of LOS (100 measurement spots) and OLOS (100 measurement spots) are considered in this measurement. In OLOS propagation case, the wireless links are semi-blocked (shadowed) by metal machines and buildings. And as mentioned in the previous part, a special case is considered that the sensor/actuator (receiver antenna of the channel sounder in our channel measurements) is placed inside the metallic body of a machine, as shown in Figure 2. The dimension of the measurement factory is 120 m×59 m×8 m, the Tx antenna height is 5 m and the Rx antenna height is 1.4m.

Table 1: Measurement Parameters.

Bandwidth Type	Frequency	Bandwidth	Power	Antenna Gain
NB	1.1GHz	0.8 MHz	33dBm	0dBi
NB	2.55GHz	0.8MHz	33dBm	0dBi
WB	5.8 GHz	8MHz	33dBm	16dBi

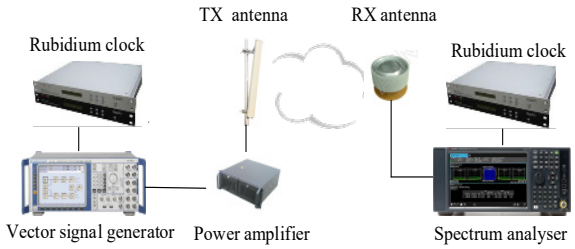
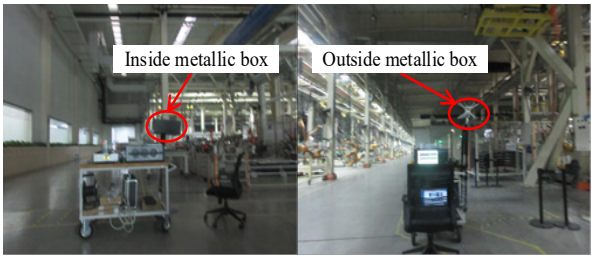


Figure 1: Channel sounding system



(a) Receiver inside (b) Receiver outside

Figure 2: Measurement setup

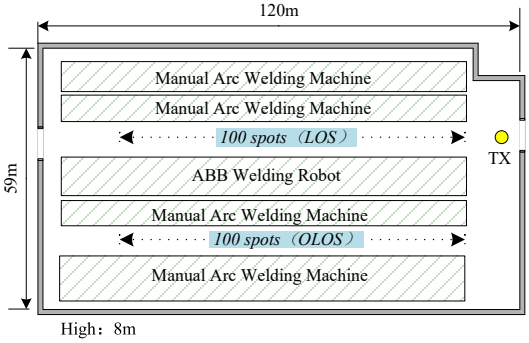


Figure 3: Environment map of the automobile factory

In the measurement, in order to extract the channel impulse response (CIRs), the frequency-domain method (OFDM signal) is

used. There are N subcarriers in one OFDM symbol with transmitted. $X[m]$ denotes the excitation signal on the m-th subcarrier. The discrete time excitation signal $x(n)$ can be given after N-point IFFT

$$x(n)=\frac{1}{N}\sum_{m=0}^{N-1}X[m]\exp\left(j2\pi\frac{mn}{N}\right) \tag{1}$$

The convolved output between the excitation signal and the time-variant CIRs denoted as a function of the discrete time variable (n) and the time delay (τ) can be expressed as

$$\begin{aligned} y(n)&=h(n,\tau)*x(n)+w(n)\\ &=\sum_{\tau=0}^{L-1}h(n,\tau)x(n-\tau)+w(n)\end{aligned} \tag{2}$$

where $w(n)$ is the additive white Gaussian noise (AWGN), L denotes the amount of MPCs. In the duration of one OFDM signal, the channel is assumed as constant, i.e., $h(n,\tau)=h(\tau)$. The received signals can be obtained as follows after removing the CP and FFT operation:

$$\begin{aligned} Y[m]&=\left(\sum_{\tau=0}^{L-1}h(\tau)e^{-j2\pi\tau k/N}\right)X[m]+W[m]\\ &=H[m]X[m]+W[m]\end{aligned} \tag{3}$$

Then the frequency domain correlation processing is performed as

$$\tilde{H}[m]=\frac{Y[m]X^{*}[m]}{|X[m]|^{2}}=H[m]+W'[m] \tag{4}$$

Finally, the extracted CIR can be obtained from (4) via the Fourier series as

$$\tilde{h}(\tau)=\sum_{n=0}^{N-1}\tilde{H}[m]e^{j2\pi\tau m/N} \tag{5}$$

Finally, the real MPCs from $\tilde{h}(\tau)$ can be extracted. And then the method in reference [12] is used to the noise threshold determination and path searching.

3. Results and Analysis of Measurement Data

3.1. Path Loss Characteristics

Path loss is an important parameter, which reflects the energy loss of electromagnetic wave in the environment. Usually, previous researchs and results show that the path loss power is mainly related to the distance between the receiver and the transmitter. And when acquiring the propagation loss power and the distance between the receiver and the transmitter, path loss prediction can be described with the empirical path loss model which can be written as

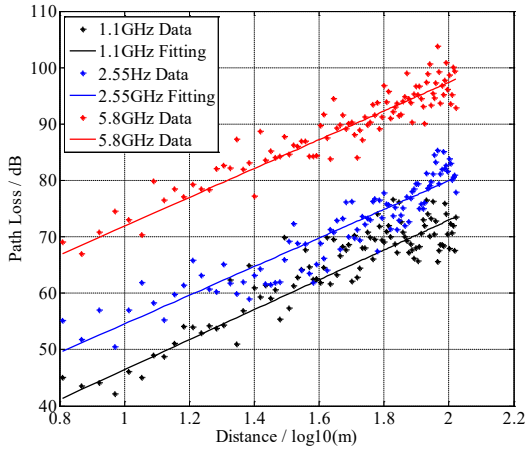
$$PL(d)=A+10\cdot n\cdot\log10(d)+X_{\sigma} \tag{6}$$

where n represents the path loss coefficient, d denotes the distance in meters between the receiver and the transmitter, X_{σ} is the shadowing term, and A represents the intercept. The least square (LS) method is adopted to the linear fitting and estimate n in the measurement data processing. In TABLE 2, the results are listed.

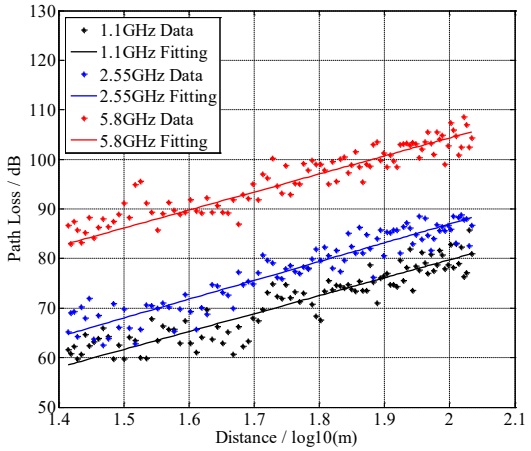
Table 2: Path Loss Parameters

Frequency	LOS		OLOS	
	n^{Ous}	n^{Ins}	n^{Ous}	n^{Ins}
1.1GHz	2.5	2.6	3.1	3.4
2.55GHz	2.4	2.5	3.4	3.7
5.8 GHz	2.6	-	3.2	-

The superscript *Ous* and *Ins* represent cases that the receiver antenna of the channel sounder is outside and inside of the metallic box respectively.



(a) Path Loss in LOS Scenario



(b) Path Loss in OLOS Scenario

Figure 4: Path Loss Exponent Extraction

Different frequency bands are measured in the same route (LOS and OLOS). The LS method is used to fit the path loss model, as shown in Figure 4, which is the situation of LOS and OLOS outside of metallic body respectively. The estimated path loss coefficient is around 2.6 in LOS, and the value in OLOS is relatively larger, around 3.2 as showed in Table 2. This is due to

the fact the propagation link is semi-blocked by the metal machines and other objects. Path loss coefficient in the industrial scenarios exhibits a higher value compared to the path loss coefficient (n=2.0) in the free space model. In contrast, path loss exponent value inside metallic box is a little bit larger than the value extracted outside. This is mainly because of the energy loss of electromagnetic waves penetrating metal boxes. Moreover, it can be found that there is no clear relationship between frequency band and path loss coefficient. In reference [13] there is also similar conclusion that path loss coefficient is independent of the frequency band. Furthermore, the results in Table 2 change significantly with the path loss coefficient acquired from different industrial wireless channel measurements summarized in [14] (Table I). It can be assumed that the path loss coefficient varies with the environment (the position, the size, the height, and density of Metallic materials or machines in the industrial factory), links configuration and frequency. In a consequence, when planning wireless communication systems in the industrial factories, it is suggested that a larger path loss coefficient should be used.

And what's more, according to the characteristics of the workshop in the industrial environment, under the same propagation conditions, if the receiver exists inside the large machine or inside the metallic body, a greater path loss coefficient should be selected compared to the external. Therefore, receiver location and environmental conditions in the workshop should be investigated before selecting the appropriate path loss coefficient.

3.2. K Factor

Two different types propagation LOS and OLOS scenarios are considered in the industrial environment. The envelope of the received signal is according with Rician distribution which varies with the probability density function written as

$$P(r) = \frac{r}{\sigma^2} \cdot \exp\left(-\frac{r^2 + s^2}{2\sigma^2}\right) \cdot I_0\left(-\frac{rs}{\sigma^2}\right) \tag{7}$$

where σ^2 represents the variance of the diffuse components, $I_0(\square)$ is the nth-order modified Bessel function of the first kind, and s^2 represents the power of the LOS path. A moment-based K-factor estimator is used to extract the K-factor [14]

$$\hat{K}_{2,4} = \frac{-2\hat{\mu}_2^2 + \hat{\mu}_4 - \hat{\mu}_2\sqrt{\hat{\mu}_2^2 - \hat{\mu}_4}}{\hat{\mu}_2^2 - \hat{\mu}_4} \tag{8}$$

where the closed forms of the even moment are $\hat{\mu}_2 = E\left[r^2\right]$ and $\hat{\mu}_4 = E\left[r^4\right]$, respectively.

In Table 3, K factor values of LOS and OLOS scenarios are shown. In LOS scenario, K factor values are about 5 dB, and due to metallic materials or equipments, the values in OLOS are little smaller than that in LOS scenario. Moreover, it is expected that when the center frequency increasing, the wavelength decreasing and more objects playing a role as reflectors, the K factor value decrease. Besides, K factor inside metallic body is smaller than outside. That is due to the metallic body generates additional penetration loss, so the dominant signal is attenuated. Meanwhile, the reflection scattering in the metallic body may be enhanced. This give rise to the occurrence of deep fades and yield small Rician factors.

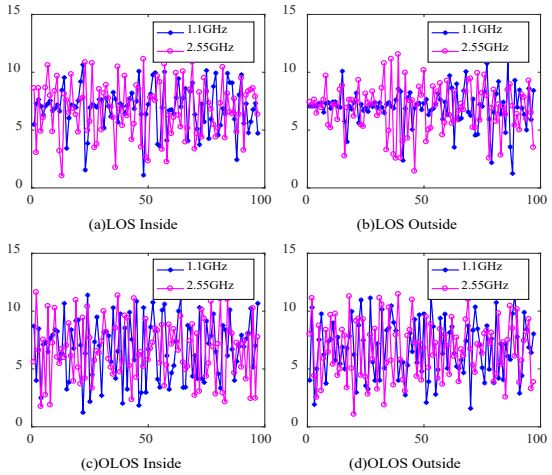


Figure 5: The change of channel envelope values

Table 3: K-Factor

Frequency	LOS (dB)		OLOS(dB)	
	K^{Ous}	K^{Ins}	K^{Ous}	K^{Ins}
1.1GHz	4.8	4.2	3.8	3.2
2.55GHz	5.2	4.5	4.6	2.8
5.8 GHz	7.3	-	4.3	-

The superscript *Ous* and *Ins* represent cases that the receiver antenna of the channel sounder is outside and inside of the metallic box respectively.

Figure 5 shows the channel envelope which represents the fluctuation degree of the receiving signals at different propagation scenarios (Figure 5 a, b, c, d) and the variances σ^2 of channel envelope in all scenarios are extracted in Table 4, which are 5.5, 3.7, 6.5, 5.6. It can be seen that when the Rx antenna is in the same environment, for example, inside of the metallic body under LOS propagation case of Figure 5(a), the envelope fluctuation of the received signals of different frequencies is basically the same. And in the same scenarios (LOS or OLOS), σ^2 inside metallic body is larger than outside. That means the fluctuation of signals envelope is more serious. This is because reflected and scattered waves in the metallic body are much more those outside of the metallic body. More MPCs give rise to significant fluctuation of the channel envelope and cause larger variance σ^2. Furthermore, it also can be found that variance σ^2 in OLOS scenarios is larger than that in LOS environments. This is due to the rich MPCs. These MPCs give rise to the severe fluctuation of the channel envelope.

Table 4: Variances σ^2 of Channel Envelope

LOS		OLOS	
Ins	Ous	Ins	Ous
5.5	3.7	6.5	5.6

3.3. Time Delay Characteristics

The time delay spread is important to digital communication system. It reflects the delay time of a multipath channel and can

represent the time dispersion characteristic parameters of the channel. The statistical results are mainly the multipaths delay spread to estimate the small-scale fading of wireless channel. The RMS delay is used to quantify the parameters of the multipath channel, the following definition of RMS delay parameters is given by

$$\sigma_\tau = \sqrt{\frac{\int_0^\infty (\tau-\bar{\tau})^2 p(\tau)d\tau}{\int_0^\infty p(\tau)d\tau}} = \sqrt{\int_0^\infty (\tau-\bar{\tau})^2 p(\tau)d\tau} \quad (9)$$

where τ is time delay, $p(\tau)$ is power delay profile (PDP), $\bar{\tau}$ is average time delay which can be expressed as

$$\bar{\tau} = \frac{\int_0^\infty \tau p(\tau)d\tau}{\int_0^\infty p(\tau)d\tau} = \int_0^\infty \tau p(\tau)d\tau \quad (10)$$

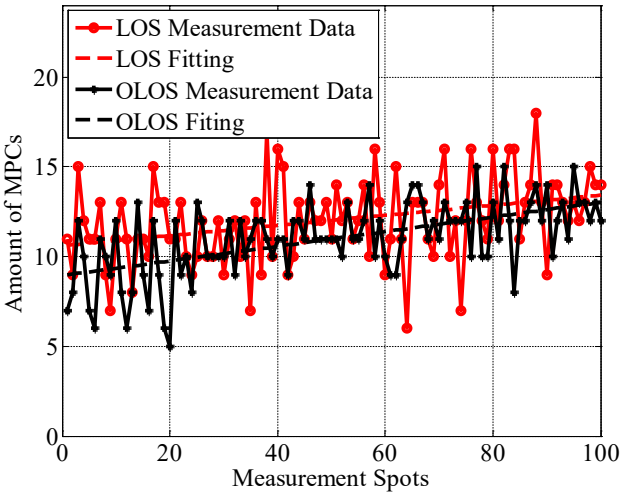


Figure 6: The amount of MPCs of 5.8GHz in LOS and OLOS

In this section, the RMS delay spread values and the number of MPCs of LOS and OLOS scenarios are researched with 8 MHz bandwidth, whether it is LOS or OLOS scenario, Out or Ins condition, there is only one main path can be obtained when the excitation signal bandwidth is 0.8 MHz (corresponding to 1.1 GHz and 2.55 GHz). And more echo waves can be obtained when the excitation signal bandwidth is 8 MHz. The main reason is that when the bandwidth is 0.8 MHz, the resolvable distance of the excitation signal is 375 m, whereas, when the bandwidth is 8 MHz, the resolvable distance is 37.5 m. So, it's much easier to make detecting MPCs when the bandwidth wider. For this reason, time delay parameters are extracted with the bandwidth in 8MHz, however, the channel envelope (not resolvable MPCs and the time delay spread) is the primary parameter extracted when the bandwidth is 0.8MHz.

Figure 6 shows the amount of MPCs of 5.8GHz in LOS and OLOS propagation conditions. It can be seen that the number of MPCs in the OLOS propagation condition is obviously more than that under the LOS propagation condition, which is mainly due to the scattered around the OLOS propagation is more abundant, and the electromagnetic waves mainly arrive at the receiving antenna by the way of reflection and diffraction. In addition, it can be seen that the number of MPCs increases slightly with the increase of distance in both LOS and OLOS propagation conditions, which is mainly due to the greater density of metal or equipment in the actual factory environment, far from the receiving end of the transmitting antenna. Meanwhile, when the bandwidth of the excitation signal is 8 MHz, the RMS results are shown in Table 5. RMS in OLOS propagation condition is larger than that in LOS, which also shows that the phenomenon of multipaths is more serious in OLOS propagation condition.

Table 5: RMS Delay Spread

Frequency	LOS	OLOS
5.8 GHz	9.37 us	15.3 us

4. Conclusions

Measurement results and analysis of the propagation fading behaviour in industrial factory environments has been presented in this paper. And two propagation types LOS and OLOS cases are considered. Particularly, special industrial propagation cases, sensors/actuators are equipped in metallic machine bodies. The results of the path loss exponent values have been reported. The estimated values are around 2.5 and 3.3 in LOS and OLOS propagation scenarios, relatively, which are larger than that of the free space model. In the research, K factor is used to described the power ratio between the direct ray and the sum of the powers of other rays. As the results shown, in LOS and OLOS scenarios, the K factor values are around 5 dB and 4 dB, respectively. And when the receiver antenna is equipped in metallic machine bodies, the path loss exponent is larger and K factor values are smaller. These are due to the metallic body generates additional penetration loss, so the dominant signal is attenuated. Further-more, channel envelope fluctuation in OLOS is more serious than LOS, and the channel envelope fluctuation inside the metallic body is more serious than outside. These are mainly because of MPCs in LOS and inside of metallic body are more abundant. What's more, in this paper time delay parameter and the number of MPCs are researched. The RMS delay spread are 9.37us and 15.3us in LOS and OLOS scenarios. The research results are important to the design of the IIoT for industrial scenarios.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research was supported in part by the Key Laboratory of Universal Wireless Communications (BUPT, KFKT-2018 105), Ministry of Education, P.R.China, Beijing Natural Science Foundation(NO. L172030), the Beijing Municipal Natural Science Foundation under Grant 4174102, and the Open Research Fund through the National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D01 and 2018D11.

References

[1] L. Liu, K. Zhang, C. Tao, K. Zhang, J. Zhang. "Channel measurements and characterizations for automobile factory environments." International

Conference on Advanced Communication Technology 2018:234-238. https://doi.org/10.23919/ICACT.2018.8323708

[2] S. Mumtaz, A. Alsohaily, Z. Pang, A. Rayes, K. F. Tsang, J. Rodriguez, "Massive Internet of Things for Industrial Applications: Addressing Wireless IIoT Connectivity Challenges and Ecosystem Fragmentation." IEEE Industrial Electronics Magazine 11.1(2017):28-33. https://doi.org/10.1109/MIE.2016.2618724

[3] T. S. Rappaport, C. D. Mcgillem. "UHF fading in factories." IEEE Journal on Selected Areas in Communications 7.1(2015):40-48..https://doi.org/10.1109/49.16842

[4] A. Miaoudakis, A. Lekkas, G. Kalivas, S. Koubias, "Radio channel characterization in industrial environments and spread spectrum modem performance." Emerging Technologies and Factory Automation, 2005. Etfa 2005. IEEE Conference on IEEE, 2006:7 pp.-93. https://doi.org/10.1109/ETFA.2005.1612506

[5] C. Oestges, D. Vanhoenacker-Janvier, B. Clerckx. "Channel Characterization of Indoor Wireless Personal Area Networks." IEEE Transactions on Antennas & Propagation 54.11(2006):3143-3150. https://doi.org/10.1109/TAP.2006.883962

[6] E. Tanghe, W. Joseph, L. Martens, H. Capoen, K. Van Herwegen, W. Vantomme. "Large-scale fading in industrial environments at wireless communication frequencies." Antennas and Propagation Society International Symposium, Conference IEEE, 2007. https://doi.org/10.1109/APS.2007.4396167

[7] M. Cheffena."Propagation Channel Characteristics of Industrial Wireless Sensor Networks." IEEE Antennas & Propagation Magazine 58.1(2016): 66-73.https://doi.org/10.1109/MAP.2015.2501227

[8] B. Holfeld, D. Wieruch, L. Raschkowski, T. Wirth, C. Pallasch, W. Herfs, C. Brecher. "Radio channel characterization at 5.85 GHz for wireless M2M communication of industrial robots." Wireless Communications and NETWORKING Conference IEEE, 2016. https://doi.org/10.1109/WCNC.2016.7564890

[9] J. Ferrer-Coll, P. ÄNgskog, J. Chilo, P. Stenumgaard. "Characterisation of highly absorbent and highly reflective radio wave propagation environments in industrial applications." Communications Iet 6.15(2013): 2404-2412.https://doi.org/10.1049/iet-com.2012.0028

[10] J. Karedal, S. Wyne, P. Almers, F. Tufvesson. "Statistical analysis of the UWB channel in an industrial environment." Vehicular Technology Conference, 2004. https://doi.org/10.1109/VETECF.2004.1399930

[11] Şeyma Tütüncü, A. Kara. "UHF propagation measurements in heavy industry." Signal Processing and Communication Application Conference IEEE, 2016.https://doi.org/10.1109/SIU.2016.7495797

[12] L. Liu, C. Tao, R. Sun, H. Chen, Z. Lin. "Non-stationary channel characterization for high-speed railway under viaduct scenarios." Science Bulletin 59.35 (2014):4988-4998.https://doi.org/10.1007/s11434-0140606 -x

[13] Cheffena M. "Propagation Channel Characteristics of Industrial Wireless" IEEE Antennas & Propagation Magazine 58.1(2016): 66-73. https://doi.org/10.1109/MAP.2015.2501227

[14] Nikolic, B. Zorica "Principles of Mobile Communication: Gordon L. Stuber; Second Edition; Kluwer Academic Publishers, Boston, 2001, 752 pages, ISBN 0-7923-7998-5." Microelectronics Journal 32.12(2001):10 50-1051.https://doi.org/10.1016/S0026-2692(01)00119-7

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 257-262 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

What Should Be Considered for Acceptance Mobile Payment: An Investigation of the Factors Affecting of the Intention to Use System Services T-Cash

Riyan Rizkyandy*, Djoko Budiyanto Setyohadi, Suyoto

Universitas Atma Jaya Yogyakarta, Yogyakarta, 55281, Indonesia

ARTICLE INFO

Article history:
Received: 01 March, 2018
Accepted: 28 March, 2018
Online: 25 April, 2018

Keywords:
Mobile Payment
E-money
T-cash
Technology Acceptance Model

ABSTRACT

E-Money mobile payments, also called digital money, are electronic payments, payment transactions using an Internet network integrated with NFC-enabled smartphones and prepaid cards. In Indonesia not only banks that issue e-money products, telecom operators from Telkomsel also issued an e-money product called T-cash. T-cash is a new innovation of electronic money presented by Telkomsel. The purpose of this study was to check the effect of responsiveness, smartness, perceived ease of use, perceived usefulness, social influence, and security against the intention to use T-cash. The data used in this study include primary and secondary data. Respondents in this study are users of T-cash products in Yogyakarta as many as 115 respondents. While the data were collected by using the questionnaire to then be analyzed using the amos analysis technique 22.0. The results of the analysis prove that two characteristics of technology, responsiveness and smartness have a significant effect on perceived usefulness. Ease of use has a significant effect on perceived usefulness. Ease of use, usefulness and security have a significant effect on intention to use. The higher the level of responsiveness, smartness, ease of use, perceived usefulness and security will also increase the use of T-cash social influence factors have no effect on intention to use.

1. Introduction

The current payment system has evolved and has evolved to abandon the old way and switch to mobile devices (m-device) now known as mobile payment (m-payment) [1]. The new payment system is the result of the development of information and communication technology in the field of economic transactions between companies and customers [2]. This payment system emerges as one way to solve certain problems related to the handling of the circulation of cash. This payment system ensures flexibility for small purchases and instant payments, enhances security and protection against fraud and other forms of crime, the emergence of e-commerce on the internet and online payments [3,4]. The current payment system should also take into account social influencing factors that will influence consumers to use mobile payment services [5,6].

E-money or electronic money is the amount of money a person keeps in an electronic medium that is officially accepted as a means of payment [7]. These payments are used for small-value transactions such as parking and public transport payments that are now beginning to use electronic cards containing e-money. And large-value transactions such as payments on the sale of goods in online shopping and transactions between other business actors [8].

In Indonesia one of the companies offering NFC based mobile payment is T-cash. T-cash is one of the services offered by Telkomsel. Telkomsel which started its business from cellular operator services then expanded its business network by offering T-cash mobile payment service. Telkomsel has obtained a license from Bank Indonesia as an electronic money service provider. T-cash can be used by all Telkomsel subscribers, either postpaid or prepaid.

Telkomsel introduced T-cash as a new generation of electronic money services as well as mobile payments. T-cash comes with innovations that will give customers and merchants an exciting new experience in making payments via mobile phones. T-cash is a product of e-money that facilitates transactions to various circles

*Corresponding Author: Riyan Rizkyandy, Universitas Atma Jaya Yogyakarta, Yogyakarta, Indonesia. Email: rizkyandyriyan@gmail.com

of society, including communities that have not been served by an authorized financial institution. T-cash was first introduced in 2015, the use of the T-cash service is easy, simply by enabling and sticking the NFC T-cash sticker to the phone [9].

With the number of e-money products in Indonesia emerging and competing, one of them is T-cash shows that technology is growing and innovation payment using e-money increasingly popular with the public. People's attention to e-money is increasingly high cause a lot of various e-money products issued by banks and other companies in Indonesia. People feel a lot of benefits and assume that using e-money products can improve or improve their social status. And some people think that new technology will only make them difficult because they are used to and comfortable with the payment of the old way. Therefore, the authors are interested to examine the issue with the title "What Should Be Considered for Acceptance Mobile Payment: An Investigation of the Factors Affecting of the Intent to Use System Services T-Cash".

Contribution in this research is social influence factor, security factor and two factor of technology characteristic that is responsiveness and smartness as an important variable that can influence the consumer to use t-cash. The contribution is important given the issue of security using electronic money greatly affects the intention to use mobile e-money payments. And in this research can also be explained that the factors responsiveness, smartness, perceived ease of use, perceived usefulness, social influence, and security are factors that may affect the intention to use e-money.

2. Hypothesis Development

2.1. Technology Characteristics

In human-computer interaction (HCI) responsiveness is an important factor because users want a responsive mobile payment system in accordance with what users expect. And using the payment system makes users look smart [10]. In [11], it is explained that two characteristics of responsiveness and smartness technologies have a significant impact on perceived usefulness.

H1: Responsiveness has a positive and significant impact on the perceived usefulness of mobile t-cash (e-money) payment system service in Yogyakarta.

H2: Smartness has a positive and significant impact on the perceived usefulness of mobile t-cash (e-money) payment system service in Yogyakarta.

2.2. TAM (Technology Acceptance Model)

In [12], it states that the Technology Acceptance Model (TAM) was developed by Davis in 1986 [13], which originated from the theory of Fishbein and Ajzen (1975) Theory of Reasoned. TAM provides a theoretical basis for knowing what factors influence the acceptance of a technology in an organization [14]. And in [15], This study examines user acceptance and use of mobile payments, focusing on mobile ticket technology applied in the context of public transport. The results show that the intention to use technology is influenced by perceived usefulness, Ease of Use and Security of that technology. In addition, perceived usefulness is simultaneously affected by ease of use.

H3: Perceived ease of use has a positive and significant impact on the perceived usefulness of mobile t-cash (e-money) payment system service in Yogyakarta.

H4: Perceived ease of use has a positive and significant impact on the intention to use mobile t-cash (e-money) mobile payment system service in Yogyakarta.

H5: Perceived usefulness has a positive and significant impact on the intention to use mobile t-cash (e-money) payment system service in Yogyakarta.

2.3. Social Influences

According to [16,17], social influencing factors are how a group or environmental factors can influence a person's behavior decisions. Social influence is a reflection of the results of communication and interaction with others so that with the occurrence of these influences can change a person's attitude or behavior. In [18], about the intent and behavior to use e-learning, those social factors are more important than usability perceptions and perceptions of ease of use, because social factors that influence students to use e-learning come from seniors as well as instructors, and lecturers are also a factor successful implementation of e-learning. In [19], explains Subjective Norm His influence can come from the views and roles of friends, peers, families, and superiors. Subjective norms have an important role in the study of the adoption of new technologies.

H6: Social influencing factors have a positive and significant impact on the intention to use mobile t-cash (e-money) payment system service in Yogyakarta.

2.4. Security

In [20], explains security is an important factor in shaping consumer confidence. The perception of security is the consumer's trust in the service user can control and maintain personal data as well as transaction data from abuse by an irresponsible person. In the study [21], found that security had a significant positive effect on online purchasing decisions. Security becomes one of the important factors because the payment transaction is done through internet network.

H7: Security has a positive and significant impact on the intention to use mobile t-cash (e-money) payment system service in Yogyakarta.

The framework underlying this research can be shown in Figure 1:

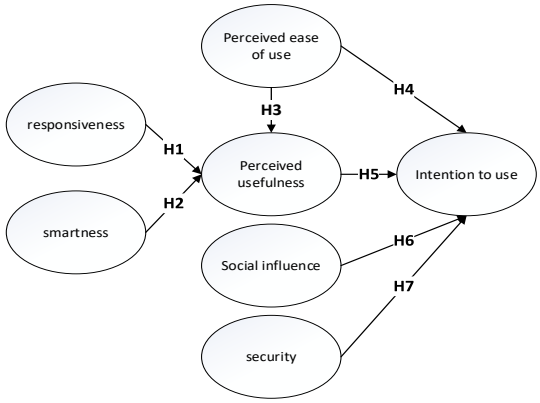


Fig 1: Research Model and Hypothesis

3. Method

The population in this research is Telkomsel subscribers who use T-cash service or who have never used T-cash respondents have unlimited nature because the number and characteristics of research respondents are not known for certain. The sampling technique used in this research is the nonprobability sampling technique. The respondents of this research are all Telkomsel subscribers who use T-cash who have access and have made T-cash mobile payment transaction in Yogyakarta.

The sample size plays an important role in the estimation and interpretation of results. In other structural methods, the sample size becomes the basis for the estimation of sampling error [22]. The sample size guidelines are 5-10 times the estimated number of parameters, the samples were taken in this study were 115 respondents (23 indicators x 5) [23,24].

In the questionnaire, there are two parts, the respondent's demographic information section, and 23 questions representing the factors in this study. Measurement of variables using Likert scale 1-5. List of questions can be seen in table 1.

Table 1: Questionnaire

Factor	Item	Question
Responsiveness (RES)	RES1	T-Cash mobile payment will offer prompt service to me.
	RES2	T-Cash mobile payment has willingness to help me.
	RES3	T-Cash mobile payment gives me individual attention.
Smartness (SMA)	SMA1	T-Cash mobile payment is intelligent.
	SMA2	T-Cash mobile payment is smart.
	SMA3	T-Cash mobile payment makes me look smart.
Perceive Ease of Use (PEOU)	PEOU1	Learning to operate an T-Cash mobile payment system would be easy for me.
	PEOU2	I would find an T-Cash mobile payment system easy to use.
	PEOU3	I think it is easy to use an T-Cash mobile payment system.
Perceive Usefulness (PU)	PU1	Using T-Cash mobile payment system in my finance related job would enable me to accomplish tasks more quickly.
	PU2	Using T-Cash mobile payment would enhance the effectiveness of my finance related job.
	PU3	I would find T-Cash mobile payment system useful in my finance job.
Social Influence (SI)	SI1	Friend's suggestions and recommendations will affect my decision to use T-Cash
	SI2	Family/relatives have influence on my decision to use T-Cash
	SI3	I will use T-Cash if my colleagues use it
	SI4	I will use T-Cash if the service is widely used by people in my community
	SI5	T-Cash will enable me to improve my social status
Security (SEC)	SEC1	I think that purchasing through a T-Cash mobile payment is secure.
	SEC2	I feel secure entering my card details for payment within the T-Cash.
	SEC3	I find T-Cash mobile payment services secure for conducting my payment transactions.
Intention to Use (ITU)	ITU1	I intend to use the T-Cash mobile payment system.
	ITU2	I predict I would continue using the T-Cash mobile payment system.
	ITU3	I plan to continue using the T-Cash mobile payment system.

Table 2 summarizes the demographic characteristics of respondents. Age is divided into 3 groups, and the most dominant age is the age group of 20-25 years of 67.0%. 51.3% were male and 48.7% were female. At the education level, there are 4 groups, namely SMP (junior high school), SMA (senior high school), S1 (undergraduate), S2 (postgraduate). which dominates the level of education S1 by 43.5%. 55.7% active transactions using T-cash and 44.3% inactive transactions using T-cash.

Table 2: Demographic of the Respondents

Demographic		Frequency	Percent%
Age	< 20 years	33	28.7
	20-25 years	75	67.0
	>25 years	5	4.3
Gender	Male	59	51.3
	Female	56	48.7
Level of Education	SMP	18	15.7
	SMA	44	38.3
	S1	50	43.5
	S2	3	2.6
Active Transaction T-cash	Yes	64	55.7
	No	51	44.3

4. Results and Discussion

4.1. Measurement Model Assessment

Assessing the goodness of fit is the primary goal in SEM to know to what extent the model is hypothesized "Fit" or matched the data sample [22,25]. The result of goodness of fit is shown in the data in table 3. Based on the Results in Table 3, it can be seen that the research model is approaching as a good fit model.

The result of CMIN / DF in this study 1,433 showed that the fit research model. The GFI value in this model is 0.823. The value close to the recommended level ≥ 0.90 shows the marginal fit

research model. The RMSEA value of this study was 0.062 with the recommended value of ≤ 0.08 this shows the fit research model. The AGFI value in this model is 0.711. The value close to the recommended level ≥ 0.80 shows the marginal fit research model. TLI value in this study is 0.941 with the recommended value of ≥ 0.90 it shows fit research model. The NFI value in this study was 0.829 closer to the recommended level ≥ 0.90 showing the marginal fit research model. Based on the overall measurement of goodness of fit above indicates that the model proposed in this research is accepted.

Table 3: Assessing the Goodness of Fit

Goodness of fit index	Cut-off value	Reserach Model	Model
Significant probability	≥ 0.05	0.000	Marginal
RMSEA	≤ 0.08	0.062	Fit
GFI	≥ 0.90	0.823	Marginal
AGFI	≥ 0.90	0.711	Marginal
CMIN/DF	≤ 2.0	1.433	Fit
TLI	≥ 0.90	0.929	Fit
CFI	≥ 0.90	0.940	Fit

4.2. Structural Model Assessment and Hypotheses Testing

Instrument Validity is evaluated based on convergent validity and discriminant validity of the indicator calculated using Amos 22.0. Convergent validity is used to determine the validity of each relationship between indicators and their latent constructs (variables). Convergent validity parameters include 3 items, the value of item questionnaire loading (≥ 0.7), the value of Average Variance Extracted (AVE) with values (≥ 0.5) and communally (≥ 0.5). (Convergent validity) is said to be high if the loading value is above 0.7. Whereas discriminant validity is seen from the loading factor of each questionnaire item with the construct representing it [22,25].

Reliable instruments are not necessarily valid, while valid instruments are generally reliable. Thus, the reliability testing of the instrument must be done because it is a requirement for testing the validity. In this connection, this study measures the reliability of data with internal consistency reliability. Internal consistent reliability testing is done by testing the instrument once, then to test the reliability of the data used indicator based on the formula Variance Extracted (AVE) and Construct Reliability (CR). And the indicator of the variable is said to be reliable if the AVE value is ≥ 0.05 and CR ≥ 0.07 [22,25]. Table 4 shows the results of Variance Extracted (AVE) and Construct Reliability (CR).

The process of testing this statistic can be seen in table 5. From the data processing, it is known that the CR value is related to showing values above 1.96 and below 0.05 for the value of P [26]. H1, H2, H3, H4, H5, and H7 support or significant (S), while H6 does not support or not significant (NS).

The findings of this study have significant implications for the intention to use mobile t-cash (e-money) payment system service in Yogyakarta. In this research use TAM model. The result of two technological characteristics of responsiveness and smartness in T-Cash mobile payments has a significant impact on perceived usefulness. The results in this study support the results of previous studies conducted in [11], which ssuggest that responsiveness and smartness have a significant impact on perceived usefulness in mobile NFC payment services in Korea. The responsiveness and smartness presented in t-cash bring a fast service to the consumer in around and small-sized form of NFC sticker. Transactions can be quickly and easily done by simply bringing the sticker to the sensor, then simply enter the pin, and the transaction is done, and make the customer look smarter by using the new model payment system presented by the mobile operator Telkomsel. With the innovation of a fast and unique shaped payment model will affect consumer interest to use t-cash (e-money).

Perceived ease of use has a significant effect on perceived usefulness, the results of which support the results of previous studies conducted in [15], which suggest that perceived ease of use has a significant impact on the perceived usefulness of mobile payments (IMMPA) for public transport. The perceived benefit of

Table 4: Standardized Item Loading, AVE, CR

Variable	Item	Factor Loading	CR	AVE
ITU	ITU1	0.756	0.8334	0.6254
	ITU2	0.826		
	ITU3	0.789		
SEC	SEC1	0.729	0.8344	0.6276
	SEC2	0.821		
	SEC3	0.823		
PU	PU1	0.769	0.8567	0.6664
	PU2	0.868		
	PU3	0.809		
PEOU	PEOU1	0.778	0.8670	0.6863
	PEOU2	0.777		
	PEOU3	0.922		
SMA	SMA1	0.835	0.8584	0.6707
	SMA2	0.717		
	SMA3	0.895		
RES	RES1	0.878	0.8474	0.6502
	RES2	0.758		
	RES3	0.778		
SI	SI1	0.825	0.8737	0.5814
	SI2	0.718		
	SI3	0.736		
	SI4	0.712		
	SI5	0.814		

Table 5: Hypothesis Test Results

	Estimate	S.E.	C.R.	P	Label	Result
PU ← RES	.372	.113	3.283	.001	par_13	S
PU ← SMA	.192	.090	2.142	.032	par_14	S
PU ← PEOU	.277	.089	3.111	.002	par_23	S
ITU ← PEOU	.293	.118	2.477	.013	par_15	S
ITU ← PU	.359	.082	4.369	***	par_16	S
ITU ← SI	-.063	.078	-.801	.423	par_21	NS
ITU ← SEC	.382	.130	2.936	.003	par_22	S

consumers towards the ease of using t-cash is to settle transactions in work and daily activities quickly. In [27], mentions if in a way perceived benefits by a consumer, the positive information from mouth to mouth can be spread widely through community forums and social media. So it is influential in increasing the intention of other consumers to use t-cash.

Perceived ease of use and perceived usefulness significantly influence intention to use. In This study support the results of previous studies conducted in [15], which in his research stated that perceived ease of use and perceived usefulness have a significant effect on intention to use mobile public transportation payment. In the use of t-cash is easy to learn in a short time to get used to the operation. Because the use of t-cash is very practical and efficient without the need to swipe like other e-money prepaid cards. Enough with tap sticker then the transaction is done. So in line with research [28], explaining that consumers will use e-commerce when a convenient technology is used easily and simply. So that easily t-cash to be used and perceived benefits in terms of effectiveness and quickly in the process of influential transactions in increasing the intention to use t-cash.

Security on T-Cash mobile payments has a significant impact on intention to use. in this study also supports the results of previous research, in [20], which in his research mentions that security has a significant effect on the intention to use mobile payment in the restaurant industry. In [29], it also reinforces that there is a positive relationship between security against intent to use mobile payment services. In transactions and their use, t-cash has a sufficiently secure and secure level of security. To activate t-cash requires confirmation from active mobile phone number to fill out the identity and give notification of each transaction history of t-cash and matching between usage of operator card of Telkomsel and t-cash card number. And t-cash will ask for a pin of every transaction made. And in [30] explaining, security is a belief in a person that the activities of transactions conducted have a high enough level of security and all about the personal information provided is guaranteed and safeguarded. With the increase of security services provided to consumers, it is influential in the intention to use t-cash.

In social influence on intention to use cannot be accepted with a p-value greater than at the 0.05 level of significance. This finding explains that social environment factors around respondents such as friends, relatives, family, and community do not support or influence the respondent to use T-cash and in the use of T-cash will not improve the user's social status. The results of this study are different from previous studies that has been done in [19] which in his research stated that the influence of social or social influence significantly influences the intention to use a mobile credit card. In [31], it also shows there is a positive relationship of social influence factors on the intention to use ICT in teaching. In [32], the results show that social influences have a significant positive impact on the intentions of use for mobile data services. In [33,34,35,36], also explains that social influences have a positive and significant influence on the use of SI. However, the results of this study are also supported by research [37], which finds that the influence of social factors on the interest of information systems utilization and social influencing factors on the use of information systems is rejected or has no effect. And the results of this study are also supported by interviews to some respondents who stated that using T-cash is encouraged by personal factors or self-interest that they feel the need to use T-cash, without any influence from friends or family.

In [38], explaining that a person's decision to choose a good and a service to be used will be influenced by several factors namely, environmental factors such as family, environment, culture, and personal factors. However, this research does not show any social influence which gives significant influence to the intention of using mobile t-cash (e-money) payment system service in Yogyakarta.

In [39], explaining factors that may affect purchasing behavior or intention to use are personal factors. Personal factors are psychologically someone who is not the same as others, therefore personal factors tend to have relatively consistent and long-lasting responses to environmental influences. In personal factors, there are characteristics: age and stage of the cycle, profession, economic situation, lifestyle, personality as well as self-concept [40]. Based on these characteristics that the differences in the results of this study can be caused by respondents who almost most of the profession are students and students. Which is the group of professions still rarely use t-cash facilities (e-money) because the profession does not have the intensity of payment transactions are too many, they choose to conduct transactions in cash or via atm. Personal factors are an important consideration in taking a decision other than environmental factors. So if the environment of the respondent has given the recommendation to use t-cash, the personal factor will become consumer consideration to determine consumer decision intention in using service of the mobile payment system of t-cash (e-money) in the future [41,42].

5. Conclusion

In this study found that, with respect to the intention to use mobile payment service system T-cash in Yogyakarta. Technological factors characteristic of responsiveness and smartness have a significant effect on perceived usefulness, perceived ease of use also shows significant relation to perceived usefulness. In perceived ease of use and perceived usefulness have a significant effect on the intention to use. The security factor also has a significant effect on the intention to use. With the conclusion of H1, H2, H3, H4, H5, and H7 are supported or accepted. Only one factor is social influence has no relationship to the intention to use. So, H6 is neither support nor rejected. That way social influencing factors will become less important than personal factors that ultimately are a consideration and consumer decisions to use them.

This study also shows that perceived usefulness, security and efficiency are important factors. The interview results concluded that for those who have never used e-money before will be a little confused how to use this service, and what benefits they get. This is due to a lack of knowledge about mobile payments [43]. So the challenge for the company is to develop and create new perceptions about consumer lifestyles in the use of T-cash, where T-cash can make the payment system easier, convenient to use and become the trend of payment, and better explain what happened and need to be done during the T-cash mobile payment process. by doing this may have an effect on the intention to use T-cash [29].

Based on the conclusions obtained from the results of the analysis in this study, researchers provide suggestions for further

research, the suggestion is expected by researchers to review research at this time and previous research by adding or using other variables not contained in this study. So it can be further known what factors can affect a person to use new technologies such as e-money products in this study. subsequent research can also use a sample with a larger number, and also increase the number of indicators of the variables for the results of research can be obtained more optimal.

Acknowledgment

I would like to thank my supervisor for his guidance on this paper and to Universitas Atma Jaya Yogyakarta, Yogyakarta, Indonesia.

References

[1] I. Mackensen, "Mobile Payments – A Strategic Forecasting Approach –," 2015.

[2] F. Liébana-Cabanillas, J. Sánchez-Fernández, and F. Muñoz-Leiva, "The moderating effect of experience in the adoption of mobile payment tools in Virtual Social Networks: The m-Payment Acceptance Model in Virtual Social Networks (MPAM-VSN)," *Int. J. Inf. Manage.*, 2014.

[3] F. José Liébana-Cabanillas Juan Sánchez-Fernández Francisco Muñoz-Leiva *et al.*, "Industrial Management & Data Systems Role of gender on acceptance of mobile payment," *Ind. Manag. Data Syst. Internet Res. Ind. Manag. &amp Data Syst. Iss Tao Zhou Ind. Manag. &amp Data Syst.*, vol. 114, no. 6, pp. 369–392, 2014.

[4] L.-C. Francisco, M.-L. Francisco, and S.-F. Juan, "Payment Systems in New Electronic Environments: Consumer Behavior in Payment Systems via SMS," 2015.

[5] K. Pousttchi and D. G. Wiedemann, "What_influences_consumers_intention_to_use_mobile," 2014.

[6] H. Khatimah and F. Halim, "The effect of attitude and its decomposed, perceived behavioral control and its decomposed and awareness on intention to use e-money mobile in Indonesia," *J. Sci. Res. Dev.*, 2016.

[7] T. J. Gerpott and K. Kornmeier, "Determinants of customer acceptance of mobile payment systems," *Int. J. Electron. Financ.*, 2009.

[8] D. Flood, T. West, and D. Wheadon, "Trends in Mobile Payments in Developing and Advanced Economies," 2013.

[9] M. S. Arifin *et al.*, "Smart Vending Machine Based on SMS Gateway For General Transactions," *IEEE*, 2017.

[10] S. Megadewandanu, Suyoto, and Pranowo, "Exploring mobile wallet adoption in Indonesia using UTAUT2: An approach from consumer perspective," in *Proceedings - 2016 2nd International Conference on Science and Technology-Computer, ICST 2016*, 2017.

[11] S. Shin, "The Effects Of Technology Readiness And Technology Acceptance On NFC Mobile Payment Services In Korea," *J. Appl. Bus. Res.*, vol. 30, no. 6.

[12] P. Moses, S. L. Wong, K. A. Bakar, and R. Mahmud, "Perceived Usefulness and Perceived Ease of Use: Antecedents of Attitude Towards Laptop Use Among Science and Mathematics Teachers in Malaysia," *Asia-Pacific Educ. Res.*, 2013.

[13] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *Source MIS Q.*, vol. 13, no. 3, pp. 319–340, 1989.

[14] R. Rauniar, G. Rawski, J. Yang, and B. Johnson, "Technology acceptance model (TAM) and social media usage: an empirical study on Facebook," *J. Enterp. Inf. Manag.*, 2014.

[15] L. Di Pietro, R. Guglielmetti Mugion, G. Mattia, M. Renzi, and M. Toni, "The Integrated Model on Mobile Payment Acceptance (IMMPA): An empirical application to public transport," *Transp. Res. Part C*, vol. 56, pp. 463–479, 2015.

[16] E. Shih, -Tse Wang, N. Pei, and -Yu Chou, "Consumer Characteristics, Social Influence, and System Factors on Online Group-Buying Repurchasing Intention," *J. Electron. Commer. Res.*, vol. 15, no. 2, 2014.

[17] G. Harvey Tanakinjal, R. MacDonell Andrias, S. Laison Sondoh Jr, and A. Asri Ag Ibrahim, "Relationship between Perceived Benefits and Social Influence towards Self- Disclosure and Behavioral Intention in Web 2.0," *Eur. J. Bus. Soc. Sci.*, vol. 1, no. 4, pp. 63–75, 2012.

[18] D. B. Setyohadi, M. Aristian, B. L. Sinaga, and N. A. A. Hamid, "Social critical factors affecting intentions and behaviours to use E-Learning: An empirical investigation using technology acceptance model," *Asian J. Sci. Res.*, 2017.

[19] G. W. H. Tan, K. B. Ooi, S. C. Chong, and T. S. Hew, "NFC mobile credit card: The next frontier of mobile payment?," *Telemat. Informatics*, 2014.

[20] C. Cobanoglu, W. Yang, A. Shatskikh, and A. Agarwal, "Hospitality Review Are Consumers Ready for Mobile Payment? An Examination of Consumer Acceptance of Mobile Payment Technology in Restaurant Industry "Are Consumers Ready for Mobile Payment? An," 2015.

[21] A. Raman and V. Annamalai, "Web Services and e-Shopping Decisions: A Study on Malaysian e-Consumer," *IJCA Spec. Issue "Wireless Inf. Networks Bus. Inf. Syst.*, 2011.

[22] T. Wijaya, *Analisis Structural Equation Modeling Menggunakan AMOS*, 5 ed. Yogyakarta: Universitas Atma Jaya Yogyakarta, 2009.

[23] J. W. Creswell, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," *Can. J. Univ. Contin. Educ.*, vol. 35, no. 2, 2009.

[24] A. Ferdinand, *Structural Equation Modelling dalam Penelitian Manajemen*. Semarang: FE UNDIP, 2002.

[25] R. B. Kline, *Principles and Practice of Structural Equation Modeling*, 3rd ED. New York: Guilford Press, 2011.

[26] J. f Hair Jr, W. C. Black, B. J. Babin, and R. E. Anderson, "Multivariate Data Analysis, 7/e," *Pearson Prentice Hall*, 2010.

[27] J. C. Roca *et al.*, "Information Management & Computer Security The importance of perceived trust, security and privacy in online trading systems'The importance of perceived trust, security and privacy in online trading systems' The importance of perceived trust, security and privacy in online trading systems," *Inf. Manag. Comput. Secur. Ind. Manag. & Data Syst. Online Inf. Rev. Iss Downloaded by MAHIDOL Univ.*, vol. 17, no. 22, pp. 96–113, 2015.

[28] M. Bonera, "The propensity of e-commerce usage: the influencing variables," *Manag. Res. Rev.*, 2011.

[29] J. Hauff, "Consumer Acceptance of Mobile Payment Services," 2011.

[30] C. L. Hsu, C. F. Wang, and J. C. C. Lin, "Investigating customer adoption behaviours in Mobile Financial Services," *Int. J. Mob. Commun.*, 2011.

[31] S. M. Salleh and K. Laxman, "Investigating the factors influencing teachers' use of ICT in teaching in Bruneian secondary schools," *Educ. Inf. Technol.*, 2013.

[32] T. Faziharudean and T. Li-Ly, "Consumers' behavioral intentions to use mobile data services in Malaysia," *African J. Bus. Manag.*, vol. 5, no. 5, pp. 1811–1821, 2011.

[33] R. L. Thompson, C. A. Higgins, and J. M. Howell, "Personal Computing: Toward a Conceptual Model of Utilization," *MIS Q.*, 1991.

[34] V. Venkatesh and F. D. Davis, "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," *Manage. Sci.*, vol. 46, no. 2, pp. 186–204, 2000.

[35] G. C. Moore and I. Benbasat, "Development of an instrument to measure the perceptions of adopting an information technology innovation," *Inf. Syst. Res.*, 1991.

[36] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward a Unified View," vol. 27, no. 3, pp. 425–478, 2003.

[37] D. P. Triadmojo, "Analysis Of Factors Affecting The Interests Of The Use Of Information Systems And Use Of Information Systems (Empirical Study On Government Of Banyuwangi)," 2016.

[38] F. Liébana-Cabanillas, J. Sánchez-Fernández, and F. Muñoz-Leiva, "Antecedents of the adoption of the new mobile payment systems: The moderating effect of age," *Comput. Human Behav.*, 2014.

[39] Kotler and Philip, *Manajemen Pemasaran*, Edisi 11. Jakarta: PT Indeks Kelompok Gramedia, 2005.

[40] Kotler, Philip, and G. Amstrong, *Principles of Marketing*. New Jersey: Person Education, 2012.

[41] S. Y. Park, M.-W. Nam, and S.-B. Cha, "University students' behavioral intention to use mobile learning: Evaluating the technology acceptance model," *Br. J. Educ. Technol.*, 2012.

[42] W. Nasri and L. Charfeddine, "Factors affecting the adoption of Internet banking in Tunisia: An integration theory of acceptance model and theory of planned behavior," *J. High Technol. Manag. Res.*, 2012.

[43] A. S. Etim, "Mobile banking and mobile money adoption for financial inclusion," *Res. Bus. Econ. JournalBold Rotman*, vol. 9, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 258-266 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Developing Students' Motivation for Learning through Practical Problems in School

Mikhail Rodionov[*,1], Zhanna Dedovets[2]

[1]*Department of Computer and Mathematical Education, Penza State University, Russia*

[2]*Department of School of Education, The University of the West Indies (UWI), Trinidad and Tobago*

A R T I C L E I N F O

Article history:
Received: 13 August, 2018
Accepted: 14 September, 2018
Online: 08 October, 2018

Keywords:
Motivation
Student
Practical problem
School
Teacher

A B S T R A C T

This paper is an extension of work originally presented in the 19th International Conference on Computer Supported Education and Information Technology. This paper identifies motivational factors that ensure the initiation and effective implementation of mathematical activity. These factors are in subordination to each other, forming a hierarchical dependence. At the heart of this hierarchy is the practical need for solving problems from real-life practice. In order to actualize this need, a number of approaches are proposed, each of which manifests itself differently at various stages of teaching mathematics at school. At the first stage, the task material is intended, in the main, only to stimulate the consideration of certain mathematical problems. It also initiates to some extent the activity procedures inherent in reality through observation and experiment. At the same time, the emphasis in teaching is on solving problems of calculation, measurement, tracing, construction, cutting, etc. At the next stage, the main emphasis is on the possibility of using a mathematical tools in the study of related disciplines. It introduces elements of mathematical modelling of real-life states and processes, which can be carried out on the basis of solving various textual problems. At the final stage, the dominant focus the vocational guidance function of practice, the mathematical knowledge for successful implementation of future professional activity. Such awareness can be provided, in particular, with the help of pseudo-real applications from the relevant professional field. In this paper the authors also provides examples supporting each stage.

1. Introduction

Successful motivation to learn mathematics cannot be achieved by a simplistic and monolithic approach, since mathematical activity is an innately multidimensional phenomenon. It requires the discriminate attention to a range of motivational factors-namely practical need, creative need, adequate and accurate language facilities, the need for proof, and aesthetic satisfaction.

These can be shown as a triangular hierarchy, mirroring the classic triangle of needs first conceptualised by A. Maslow Figure 1.

At the base of this triangle lies the practical need for solving pressing problems within the realm of day to day human activity. The solution of each of these tasks presupposes the need to generalize the empirical material accumulated in the course of each practical activity, a generalised extraction of certain characteristics common to many objects and phenomena. This is not possible without a definitive leap in the development of human thinking, which marks the emergence of a creative need for the discovery of new facts and patterns that do not belong to the sphere of immediate utility.

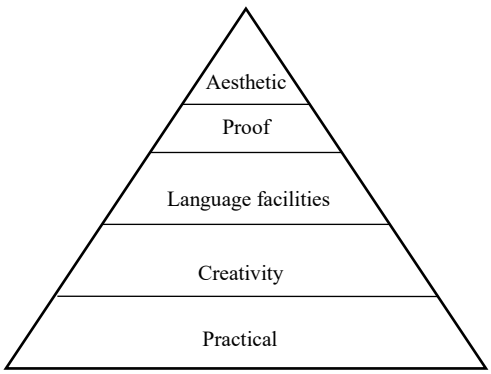


Figure 1 Hierarchical model of mathematical activity needs

[*] Mikhail Rodionov & Email: do7tor@gmail.com

The historical accumulation of schemes and decision algorithms available for an elementary study of problems necessitated the clarification of terminology and the development of more precise and, at the same time, more symbolic means of presenting these problems than could be achieved using natural language. This encouraged the logical organization of mathematical knowledge in the form of a harmonious deductive theory. At the same time, there was a reassessment of the means for checking the discovered regularities, which, in turn, led to the emergence of the need for their proof using system predetermined postulates and axioms.

Further improvement of mathematical education was achieved through, in particular, attempts to minimize the set of initial assumptions and interpretations of various mathematical theories in such a way that the derivation of new regularities would be seen as elegantly simple, unforced and economical. These attempts were a manifestation of a new aesthetic need to create the perfect expression of the hitherto ad hoc description of motivational factors.

The transition to more advanced levels of the described hierarchy does not invalidate the discrete functioning of the "lower order" needs. On the contrary, these needs are also improved and enriched by entering into productive interaction with the "higher" motivational factors.

These factors are simultaneously present in any mathematical activity. This is especially true for creative needs, which are ideally realized at the initial stage of posing the problem, and then at the subsequent stages of choosing the path of the solution and its justification. The need for proof is clearly expressed at the stage of justifying the decision, and the need for effective language facilities is needed at the stage of formalizing the initial problem. Practical need is especially evident at the stage of problem selection and at the stage of correlating the already obtained result with the limitations provided by the specifics of this problem. Finally, the aesthetic need can potentially be actualized at all these stages, even after solving the initial real situation, by seeking the possibilities of expanding derived and valid mathematical regularity.

The presence of a particular need of the researcher is only a necessary condition for the productive course of mathematical activity. What is also required is the possibility of "objectifying" and internalising this need into consciously understood motivated activity. This in turn increases practical capacity which in turn can increase conceptual understanding in a virtuous circle.

2. Practice as a source of mathematical creativity

The practical need for the realization of mathematical activity stems from the specific nature of the subject of mathematics as distinct from other scientific disciplines and is embodied in the nature of the interaction of this discipline with reality and public practice. Most scientific disciplines clearly relate to some form of the movement of matter, or to the sphere of individual and social practice, which they study with the help of a variety of methods (including mathematical ones), but, as a rule, remain within their subject area. Mathematics does not set as its immediate goal the analysis of any specific phenomena and processes, but as its raison d' etre specifically distinguishes the quantitative relations and spatial forms inherent in all subjects and phenomena without exception, and considers them as the purpose and object of its research. Such research is always carried out on the basis of formal approaches "potentially admitting the most diverse material incarnations, and consequently, applications". In other words, mathematics can be seen as a "universally applicable scientific method", a kind of generalized "working scheme" for research, description and cognition of nature [1, 2].

This approach has always been associated by specialists in the field of the methodology of science with the general philosophical question of the reasons for the effective value and universality of mathematical theories in solving problems arising in the course of human interaction with the environment. Regular dependencies, expressed by mathematical laws, can be embodied in the intrinsic nature of the objective external world, and we only discover these dependencies through experience and experiment. At the same time, the deductive, speculative nature of mathematical knowledge can reflect its relative isolation from other spheres of human activity. In this latter case the applicability of mathematics to practice can seem bafflingly remote. The first perspective brings to the forefront the external source of the moving forces of mathematics, while the second focuses its attention on the internal needs of the development of this science and its systemic representation.

In actual mathematical activity, both tendencies are almost never presented in isolation. The practice of pure mathematics can subsequently find important practical applications, while results initially assumed to be solely applied do not in fact find any practical applications. So, for example, mathematical logic, which previously mattered only to the persons dealing with the problems of justification, have in recent decades begun to be regarded as an applied science closely connected with computational mathematics. The set-theoretical concept, traditionally perceived as the theoretical foundation of all modern mathematics, is now directly used for the analysis of phenomena of the most diverse sciences - from biology to linguistics [3].

At the same time, it is possible to give examples of so called applied theories, such as the "paint brush theory", which have no value from either a practical or a theoretical point of view. An interesting example of this nature is suggested by J. Stuart. This author tells about a man who, from general mathematical considerations, derived a very complex formula, filled with constants e, c, h for calculating the radius of the universe. And it was only after many years that scientists decided to obtain by using this formula a specific value of the radius. It turned out that it is equal to 10 centimeters [1].

R. Courant and G. Robbins emphasized that discoveries which simultaneously meet theoretical and practical needs are of particular importance for the further improvement of mathematical knowledge [4]. Thus, for example, the transition from natural to rational numbers served a theoretical need to remove restrictions on the performance of the corresponding arithmetic operations, and also the practical necessity for numbers suitable for the results of measurements of quantities.

In actual mathematical activity, external and internal stimuli, as a rule, are difficult to distinguish. Practice itself often influences mathematics, not only because it immediately requires extensive

and profound mathematical knowledge, but also because it can create insights into the unraveling of the mysteries of nature and of the properties and activities of otherwise mysterious configurations. In addition, it is the development of practice certainty, and the need for unambiguous, consistent and correct interpretation of certain natural patterns which has been one of the main motives for developing relatively uniform principles for the systematization and organization of mathematical knowledge, which in turn led to their qualitative transformation and to the formulation of the deductive method.

Let us consider in more detail the manifestation of the stimulating role of practice in relation to mathematical science. The most transparent role is manifested when developing special mathematical methods for solving specific problems that arise in peoples' real life activities. Among such problems in the early stages of the development of mathematics are the problems of land surveying, the calculation of the volume of vessels, the practical calculation, the calculation of time and the prediction of natural phenomena. Somewhat later, practical changes and developments in, for example, trade, construction and agriculture, combined with new challenges in areas such as astronomy, geography, mechanics, and optics-accelerated the predominant development of computational methods.

A general description of the realization of the relationship between mathematics and practice in similar cases presupposes a sequence of steps, as reflected in Figure 2 below.

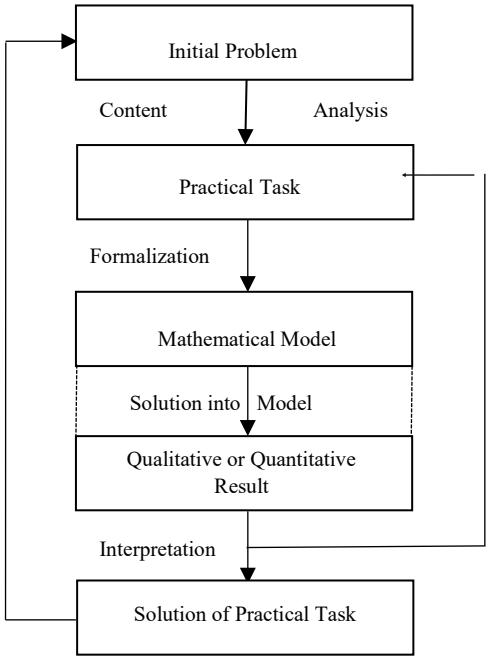


Figure 2 The realization of the relationship between mathematics and practice

In accordance with the above Figure 2, a person's activity in the investigation of one or another side of the real world begins with a meaningful analysis of the initial general problem situation, leading to the formulation of a more specific concrete practical problem. This analysis includes a refinement of the meaning of initially insufficiently defined input parameters; checking the completeness of the initial information and introducing, if necessary, missing data; choosing the range of accuracy and numerical data corresponding to the nature of the problem, and making a preliminary assessment of the value of the problem being solved in relation to the initial problem situation [5].

At the next stage, a mathematical model of the problem is constructed, reproducing the features of the structure and properties of the original in the language of mathematical terms and symbols. Next, the model is transformed, with the aim of obtaining specific numerical values. After correlating these values with the initial situation, a conclusion is made about the plausibility of a given result and the possibility of using the method of obtaining it in solving problems of a similar nature. As an example of this possibility, one can indicate the application in electromagnetism and optics of mathematical methods originally intended for the theory of elasticity [2].

The dominant motivational factor in any considered case is the urgent need to obtain a solution of this particular practical problem irrespective of other possible areas of application of the constructed mathematical model. In such a situation, the effect of this motive essentially ends with the result. A relatively recent example of such a "direct" impact of practice on the development of mathematics is that of linear programming, which arose on the basis of a number of particular problems (optimization of material consumption, organization of transport, etc.) [3, 5, 6, 7].

It should be noted that in the course of the historical development of mathematics, the limiting scope imposed by the way a task was framed has often had a negative effect on the improvement of mathematical knowledge. For example, ancient scientists, in solving the problem of "incommensurable" values, could not overcome the Pythagorean numerical traditions and build mathematics on the basis of pure axiomatic geometry. This resulted, in the words of R. Courant, in one of the "strange wanderings in the history of science", which for two millennia delayed the "evolution of the idea of number and alphabetic calculus" [1].

A more positive effect of practice on the development of mathematics occurs when mathematical thinking goes far beyond what the posed practical task directly demands, passing successively through a series of steps of abstraction of the concepts and methods from their originally concrete and material prototypes. At the same time, the increasing abstracted purity of mathematical theories simultaneously increases their applicability, so that the range of problems under consideration is broadened and generalized, which, in turn, facilitates the transition of any given theory to a new stage of abstraction. This regularity can be represented schematically in the form of a system of "embedded" plane figures embodying the corresponding levels of abstraction of the practical problems (Figure 3).

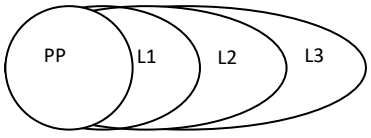


Figure 3 The levels of abstraction

We will consider this Figure 3 on the basis of one of the key directions in the development of mathematical science.

In solving the problem of the listing of any objects, each of the individual properties that are available for each of these objects can begin to play the role of a "unit". A search of these "units" leads to the establishment of a certain sequential order and the emergence of the concept of the ordinal, and then of the quantitative number. The implementation of the simplest operations on numbers convinces us of the need to extend the notion of a number, which has as its final result the creation of a single logical concept of arithmetic.

At the next level of abstraction, it is appropriate to introduce symbols, which imply any quantitative numbers. Thus, we move from arithmetic to algebra and then to mathematical analysis, which unlike algebra studies variables not only in discrete but also in continuous form [8].

The rapid development of theoretical natural science in the late 19th century meant mathematics had also to develop to a new, higher stage of abstraction. In addition to matters of quantitative symbolic content, the mathematical operations themselves are considered to be variables.

The highest level of abstraction of modern mathematics, as already noted above, does not mean its separation from practice. On the contrary, abstract mathematics, operating with such abstract concepts as a group, a set, an abstract space, has managed to explain and express the most complex processes and phenomena of 'reality'. The application of mathematics to practice is at this level generally realized directly in indirect conjunction with natural science disciplines. As an example, the practice and development of topology produced changes in differential geometry, enriched the theory of relativity and, through the theory of critical points in the calculus of variations stimulated the development of homological algebra, and, through the theory of sheaves and cohomology, algebraic geometry. Thus topology has moved from an esoteric category of mathematics to one of its central unifying and basic facets [3].

During the transition through levels of abstraction, mathematics extracts additional information hitherto implicit in its structures, while providing itself with the possibility of continuing applicability to pressing practical demands and also strengthening its internal potential. It should be noted that the initial practical need does not disappear without a trace in the course of its immediate satisfaction. It implicitly participates in the construction of a chain of internal motives for the improvement of mathematical knowledge, reinforced from time to time by direct external impulses. These impulses are due to the classical and ongoing tensions within science between the newly obtained results of experience and observations and the basic laws underlying the corresponding theoretical concept. The solution of these contradictions assumes either the search for new, still unknown methods, or the amendment of the already adopted basic provisions with the aim of increasing the accuracy of their wording through the use of second, third, etc. degrees of approximation. For example, in the early 20th century mathematicians realized that the theory of quadratic forms developed by D. Hilbert was not entirely suitable for solving the problems of rapidly developing quantum mechanics. Awareness of this fact served as a stimulus for J. von Neumann to improve this theory. By refusing the traditional "binding" of a quadratic form to a specific algebraic record, Neumann was able to produce a more general definition which avoided the limitations of the Hilbert approach. This improved theory of quadratic forms was able to provide answers to the very real and specific demands of modern physics [3, 5].

Another possibility of the stimulating effect of practice on mathematics is the transfer of concepts, representations and modes of activity which are characteristic of natural and humanitarian disciplines, to the field of "pure" mathematical science. This possibility stems either from the absence of concepts in mathematics, yet in which language one can describe certain real processes and phenomena, or from the presence of certain discrepancies in the interpretations of "related" objects of study in mathematics and corresponding fields of human knowledge. A change in the viewpoint of mathematical methods makes it possible, in a number of cases, to stimulate the development of a corresponding section of mathematical content. For example, the use of a rigorous definition of the limiting transition in the sense of Augustin-Louis Cauchy in the study of real processes seems at first sight to be impossible, since the consideration of any physical quantities reduced "beyond some reasonable boundaries" is completely meaningless. In connection with this, in physics, the so-called "practical infinitesimal quantities" are considered, and treated as actually infinitesimal. The apparent contradiction is resolved on the basis of a scientifically grounded possibility of using the definition of a limit transition not only for infinite continuous processes, but also for particular cases on bounded sets. Despite this possibility, the practical treatment of infinitesimal quantities has found its application in the concept of nonstandard analysis that goes back to Leibniz, which in a number of cases substantially simplifies the classical exposition of it [4, 6, 9].

Among the existing "channels of influence" of practice on mathematical activity, one more should be mentioned. As we know, practice is the source of plausible reasoning based on intuition, experiment, analogy and constructive induction. These arguments, unlike the proofs, do not provide the reliability of the mathematical theory. However, in them, according to R. Courant and G. Robbins, can be found the real essence of any mathematical discovery, even if it belongs to its most abstract areas [1]. Many examples of how plausible reasoning can lead to the discovery of certain mathematical regularities are given in the well-known work of G. Polya [10]. It is important to note that this author directly relates the application of these arguments to motives that are starting points for the promotion of relevant hypotheses and their proofs.

3. The motivational role of practice for mathematical education in school

The assimilation of a mathematical theory requires consideration of empirical concepts directly related to practical activity. These concepts can also be fully understood, refined and used in practice as meaningful interpretations of abstract theoretical concepts. It is necessary in this respect to know about the stages in the assimilation of educational material. At the first stage, following observation and experiment, a basis of understanding of a fragment of the mathematical theory is formed. Then, in the course of understanding the whole system of empirical concepts and interrelations between them, knowledge ascends to the theoretical level. Finally, mathematical concepts and methods

of activity produce some concrete, meaningful interpretations that make it possible to intensify the students' desire to apply the acquired theoretical information in practice. This process must of course take into account the overall (and especially age related) context of school based mathematical education.

At the earliest stage, practical activity mainly performs a stimulating function in the study of mathematical content, resulting from the tendency inherent in the child's initial consciousness to "cling" to specific and directly "tangible" facts and situations [4, 5]. At the same time, the study of mathematics is a process of empirical cognition, in which observation and experiment (calculation, measurement, drawing, construction, etc.) play a major role. The main motivational factor here is the inherent desire of any person to connect the acquired material with their own life and practical experience.

In the following stages of training, the stimulating role of practice, although it ceases to be dominant, nevertheless retains its role as an important means of motivating consideration of a fragment of content and the sparking of initial interest. At the same time, the mathematical fact begins to act not so much as a straightforward step of generalization of empirical material realized by direct teacher guidance, but rather as a result of solving a purely mathematical problem, specially formulated during the consideration of the corresponding practical problem. The possibilities for such work are significantly increased when connecting the material of related subjects, especially physics. Thus, for example, the study of the propagation of light reflected from a specially placed mirror in its path leads to the question of choosing the path by which light travels the shortest distance. An attempt to answer this question leads to the formulation of a purely geometric problem. The solution of this problem is realized, as is well known, on the basis of the symmetry method. Thus, the physical problem appears here as a carrier of motivation for the application of the symmetry method in solving geometric problems.

In another case, observations of sunset followed by plotting the time dependence of the moment of sunset from the date of observation can be used to motivate the introduction of the trigonometric function y=sin x and to reveal some of its features, making it easier for schoolchildren to assimilate such concepts as the domain of function, monotonicity, zeros and the periodicity of the functions [11]. Here, empirical activity does not end with the very introduction of the concept of trigonometric function, but "permeates" the entire content of the topic, periodically providing additional support to the motivational mechanism involved in the initial situation. The main didactic condition for the effective implementation of such situations is the provision during the educational process of the possibility of their detection, awareness and successful resolution.

The participation of schoolchildren in the process of the emergence of new concepts by abstracting and generalizing the phenomena of the real world, significant though it is, by no means exhausts the stimulating potential of practice in the study of mathematical material. In particular, an essential role in the realization of this potential is played by an emerging awareness that knowing a particular fact for solving an important problem, whilst complete in itself, can also lead to the further development of the problem or to proving the theorem. The proposed problem or theorem should always be linked to students' previous, empirically based experience.

For example, at the primary level students meet challenges to find the sum of certain numbers in a number sequence:

$$1+2+3+\ldots+99+100$$

To find the sum students change the order of numbers in the sequence and group them accordingly:

$$\frac{(1+100)+(2+99)+\ldots+(100+1)}{2} = \frac{101\times100}{2} = 5050$$

At the senior level, students can apply this fact using geometric material. Here students must find out the number of lines connecting a certain number of points. After some practical actions with two, three, four points, students determine how many lines can be drawn through 100 points. As a result of discussions, students establish that the 100th point can be connected with another point by 99 lines, the 99th point by 98 lines, the 98th point by 97 lines, and so on. This allows a transition from a geometrical problem to an algebraic problem. Further analysis helps to find the number of lines passing through n points. To solve this problem students need to find the sum of the first n natural numbers, i.e. finding the sum of arithmetic progression corresponding to a number sequence.

In the given example, students were stimulated to learn the progression from manipulating geometric objects to formulating a geometrical problems and then converting it into algebraic problem. The use of geometric material gives pupils a subjective feeling of novelty. At the same time it allows them to engage in practical activities.

At the next stage of mathematical preparation, the motivational role of practice is expressed in the realization of its worldview function. Such an implementation is possible through the demonstration of the application of the studied mathematical content in related courses and other school disciplines and consideration of the history of the emergence and evolution of scientific concepts and methods. It also develops familiarity with the elements of mathematical modeling of real states and processes, underlying the mastery of applied mathematical ideology [7, 11, 13]. In addition, an understanding of the role of mathematical knowledge as an important component of human culture becomes one of the leading motivational factors, creating a conscious desire by students to use the acquired material in related subjects and real life practice.

Textual problems are a traditional means of demonstrating the practical importance of mathematics as it is studied in school. In solving them, students become acquainted with the basic l method of cognition of reality via the concept of mathematical modeling of the initial real life situation. They learn to choose which model and how to construct and apply it. They thus also learn how to analyse and interpret the quantitative, graphical or qualitative results they obtain.

Textual problems form part of the very first mathematics lessons, thereby implicitly preparing junior students for a future explicit understanding of the concept of modeling. As a rule, here they act as 'pseudo-real 'problems, presupposing exactly as much

data as necessary for the solution, and providing for an unambiguous and exact answer. The process of solving such problems is, as it were, "monologic". It does not require the distinct special definitions and refinements which characterise applied problems of a tangibly external nature [12, 14]. Accordingly, the motivational effect of such tasks is mainly determined by the artificial actualization of problematic situations, often of an entertaining nature, the plot of which to some extent correlates with the real non-mathematical experience of students. The presence of such an effect implies compliance with the following minimum set of requirements.

1. The given tasks must correspond to the real characteristics of the non-mathematical objects described.

2. The task plot should be relatively "close in spirit" to the student, reflecting the significant aspects of his or her experience.

3. The set of mathematical tools mastered by schoolchildren should be sufficient to solve the problem at a level corresponding to this stage of mathematical preparation.

As student mathematical understanding develops, so the possibilities of realizing the worldview function of practice are substantially increased. This is primarily because the set of the subject tools (the types of equations studied, the functional dependencies, algebraic expressions, etc.) are substantially enriched by each of them having the potential to be understood as a mathematical model of some real phenomenon or process. The study of related disciplines (physics, chemistry, geography and others) provides the school mathematics curriculum with a significant number of practical applications and motivations. Mastery of the intrinsic conceptual qualities of mathematical modeling of real processes, studied in related fields of knowledge, provides an understanding of the general possibility of applying mathematical knowledge and skills, and thereby develops enthusiasm for such application. At the same time it becomes the basis for the formation of educationally cognitive motivation both in relation to mathematics itself and in relation to other affected disciplines [15].

It is important to identify a number of conditions for the effective implementation of this ‘worldview‘ function of practice, which apply to all stages of the school based mathematics curriculum.

1. Maximum correlation between the mathematics curriculum and other school disciplines, with close consideration of real life practice. This correlation depends on careful planning and timing, a unity of approach to the formation of concepts common to these courses; consistency of terminology, notation, systems of units of measurement, and also the correspondence of operational structures to the solution of typical problems.

2. Purposeful development of mathematical intuition, implying the introduction of features specific to applied activity in the teaching of mathematics [15, 16]. This includes the basic skills and techniques used in solving practical problems (selection of the necessary data, estimation of the result, methods of approximate calculations, etc.), and the cohesive consistency of the steps characteristic of applied activity (analysis of the real situation, formulation of the problem, choice and construction of its mathematical model, interpretation of the real meaning of the result).

3. Clear demonstration of the origin and development of mathematical concepts and methods because of related knowledge and real life needs and experience.

All this is best achieved within a common ethos and a culture of dialogue within each and across all parts of curriculum. This produces mutually enriching learning within and across disciplines. It is also a matter of positive and sensitive approach. Factors such as emotional tone, confidence in the cognitive abilities of the interlocutor, and mutual support are important examples. Tasks of a practical nature provide great opportunities for developing an iterative dialogue of learning because of their initial empirical uncertainty. The process brings clarity, relevance and priority. Students create a virtuous circle of learning and motivation.

As an example, let us consider a practical problem depicting a real situation associated with the construction of a cottage.

Problem 1.

The cottage has width $a = 5$ m and length $b = 8$ m. What is the size of the mansard if the distance from the attic flooring to the top of the mansard is $h = 3$ m?

After discussing the meaning of the terms related to the given situation (mansard, attic flooring, top of mansard) students find out that the length of the mansard is the length of the cottage. To find other sizes students analyze the situation and answer the following questions.

1) What geometric shapes correspond to the images of the façade of the cottage and the cross section of the mansard? The answer is a rectangle inscribed in a triangle.

2) When drawing variants of the specified geometrical configuration, when does the mansard becomes the most spacious? Check your answer by calculation. The answer is the rectangle sides are slightly different from each other.

3) How to mathematically characterize the most rational form of the cross section of a mansard? Its area is maximum.

4) What problem can you formulate using this fact? You need to find the maximum cross-section of the mansard.

The problem requirements are not imposed on students by the teacher. They are requirements which develop as an intellectual acquisition as a result of discussion.

Next, some preliminary quantitative estimation of an expected result should be carried out. The estimated value of this result should be determined. Rough approximation shows that the result is equal to the area of the cross section of the mansard which has a triangular shape (15 m^2). More precisely, this value is determined through a drawing using students’ visual and intuitive reasoning. This value equals half the cross-sectional area of the mansard approximately (7.5 m^2). This process allows students to control the progress of the search.

At the next stage of solving the problem, students’ constructive team work continues. They build a mathematical model. Using the

geometrical drawing students formulate a mathematical analog of the practical problem:

Find the sides of a rectangle inscribed in the biggest triangle. The triangle has base *a* and height *h*, (Figure 4). Introduce the notation: *LD=y; CE= h; DM=LG=x; NK=a*

Then students construct an analytical model of this practical problem using similarity of the triangles *NCK* and *LCD*:

$$S = \frac{a}{h}(h - x) \times x,$$

Where *S* is the area of the rectangle *LDMG.*

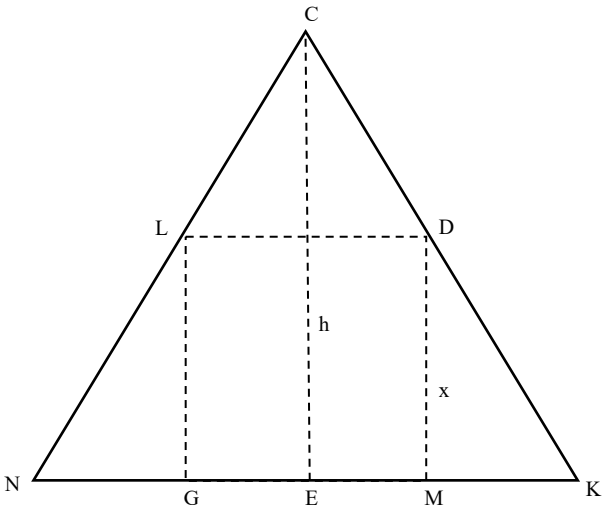


Figure 4 The mansard

An investigation of the relationship between the area of the rectangle and the second and the third factors (*a/h* constant) leads to the mathematical question: at what value of x is the product of *(h–x)x a* maximum *(x<h)?* This product can be presented in the form $-x^2 + hx$ and corresponds to the quadratic function

$$f(x) = -x^2 + hx. \quad (1)$$

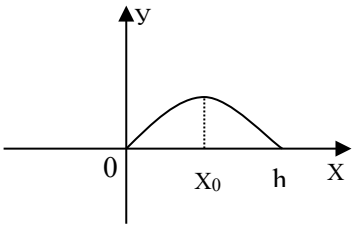


Figure 5 Graphical model

In this case only a graphical method is possible. Students construct a new graphical model of this practical problem. This model is a part of the parabola which opens downward (Figure 5). It leads to the next mathematical question. At what argument value is the function $f(x) = -x^2+hx$ a maximum?

Clearly the argument value corresponds to the abscissa of the parabola vertex. From the concept of symmetry or the formula this value is $x_0 = \frac{h}{2} = 1{,}5$

We come to different interpretations of the results.

1. An initial situation.

The maximum cross-section of the mansard is when the height of the mansard is equal to half the distance from the attic flooring to the top, i.e. *1.5 m*. The area of the cross-section is

$$S = \frac{a}{h} \times (h - x) \times x = 3{,}75m^2 (S < 7.5m^2)$$

Maximum mansard capacity is $V = S \times b = 30m^3$

2. A geometrical model.

The side of the rectangle inscribed in the biggest triangle is the middle line of this triangle.

3. An analytical model.

The product of two positive factors has maximum value if they are equal (the sum of these factors is constant).

4. A consequence.

Among all rectangles with the same perimeter, a square has the maximum area.

5. A graphical model.

The relationship between an area of a rectangle inscribed in a triangle and its side is a part of a parabola. The abscissa at the top of the parabola corresponds to the side when an area is maximum. The abscissa of the points of intersection of the parabolas and the coordinate axes corresponds to the side when an area is minimal, etc.

As a result of this work a teachers inform the students that the practical problem considered is one from a wide range of problems where one needs to find maximum and minimum. These problems are of great practical importance. It is useful to give students homework: choose similar problems from their life experience.

This example illustrates all the major stages of solving practical problems.

Among these stages the key element is the choice of the basic model of the studied situation. Practice shows that students have challenges in choosing this model without prompted training activity organized by the teacher. For this to be effective, the teacher must help students develop particular skills. For example, students must establish the similarity between various explanations, to estimate the outcome of each of them in a specific situation, whilst evaluating different approaches. This kind of work can be organized through specially chosen sets of practical problems.

Problem 2

What kind of measuring instruments do you need to determine the area of the steel plate in the form of an equilateral triangle measuring a cm?

Most students propose to use the formula: $S = \frac{a^2\sqrt{3}}{2}$ for the area of the triangle. It is sufficient to measure the side of a triangle with a ruler. The content of the school physics course allows them to determine the area through the volume and density of the material of the steel plate: $S = \frac{V}{n} = \frac{m}{p \times h}$.

Students should weigh this steel plate and find the appropriate value of density in a table. After this activity students get identical results using different formulas and different measuring instruments. The relationship between mathematical and physical reasoning then becomes clear to them.

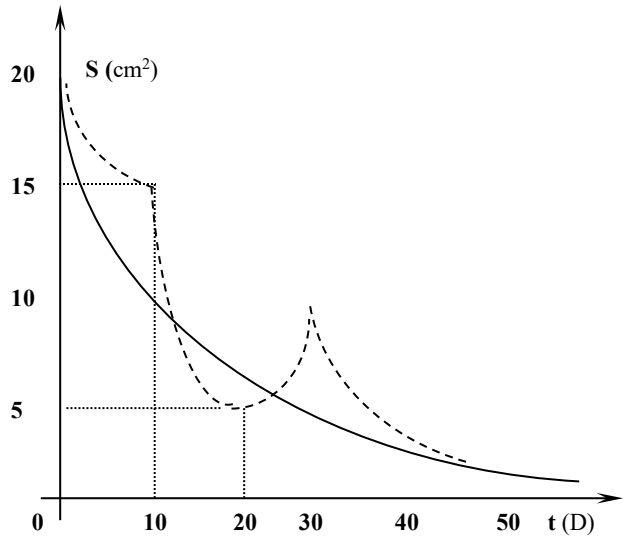
Students can see that those two approaches produce an approximate measurement. Then they have to explain the equality: $\frac{m}{p \times h} = \frac{a^2\sqrt{3}}{2}$. The obtained "motivational impulse" can be used by the teacher to clarify the range of possibilities of the methods used. A teacher gives students steel plates of different forms (a circle, a rectangle, ellipse etc.) and discusses the benefit of different approaches to solve practical problems. The practical approach is optimal for any given activity. Students may see plates of various shapes. Also, they can use the weighing method to determine the characteristics of geometric shapes. This method was successfully applied by Archimedes for the volume of a sphere. During this activity students recognize how a practical method relates to a mathematical method. This method allows students to consider from a given problem how to find areas and volumes of geometrical shapes in general. As a result of this activity students will have long-term motivation in learning future mathematical analysis. It is important to note that in the above (and indeed other) practical problems mathematical activity is not a closed and separated structure; it is a natural component of the universal system of knowledge about the world [9].

4. Students' orientation in the modern world through solving practical problems

At the last stage of teaching, the role of practice in providing mathematical orientation becomes dominant. Practice gives students a new motivation and understanding of how mathematical skills are needed to fully participate in the modern world and for the successful implementation of future professional activity. However, this is not always easy. One challenge is that many modern scientific fields operate in such a way that their types of models do not smoothly integrate with the traditional mathematics curriculum. More subtle approaches might be then needed to impart an understanding of the practical application of mathematical knowledge to future professional development and in relation to environmental and other extracurricular activities.

Let us consider an example from the field of medicine. In studying the exponential function in the school, students' attention is drawn to the traditional formula expressing the laws of growth ($y=e^{kx}$) and decay ($y=e^{-kx}$). At the beginning of the 20th century American scientists revealed the law related to the latter formula. It reflects the approximate dependence of the area of protracting wounds from the time when the wound becomes sterile. This dependence can be traced with a special device, a planimeter. A planimeter is used for approximate measurement of the surface area bounded by lines. The perfect curve of wound healing is described by the formula: $S=S1e–kt$, where $S1$ is the wound area at the initial time. The perfect curve is also called the prediction curve. A prediction curve is compared with an actual curve (Figure 6). The wound is infected if the observed wound area is larger than the area defined by the perfect curve. If a wound heals faster than the perfect curve shows, this indicates the appearance of secondary ulcers. A wound is healing well if the prediction curve is the same as the actual curve.

Such examples, as shown in our monitoring of the teaching-learning process, significantly enhance the general attractiveness of mathematics for students and across disciplines.



Based on the previous discussion, the motivational characteristics of practical problems can be presented in a systematic form in Table I.

Table 1: Systematic form of practical problems

Stage	Role of practical problem	Motivating factor	Objective	Instructional techniques	Students' activity
1	Stimulate students' interest	Intention to link learning material with their own life experiences	Deriving of mathematical relations; understanding concepts, theorems, algorithms and their application	Organization of empirical support for students' activity	Using students' life experience, applying it to solve non-standard word problems
2	Develop students' world outlook	Intention to apply mathematical material which was studied in other related school disciplines and real-life practice	Demonstration of practical application of mathematics in various branches of human knowledge	Forming of modeling method	Using plausible reasoning to solve practical problems related to other school disciplines
3	Students' orientation in the modern world	Intention to apply mathematical tools for association to their environment	Awareness of importance of mathematical knowledge and skills for further education and profession	Understanding of the features of mathematical models and their use for practical activity	The development of research with the involvement of the appropriate mathematical tools

This table demonstrates the implementation of the motivational role of practice and instructional techniques related to students' activity. The majority of the characteristics reflected in Table I are not tied to a particular stage of the teaching process. Mathematical knowledge and skills are enriched when students move to the next stage, which reinforces the stability and depth of their motivation.

5. Conclusion

The study of the nature of mathematical activity identifies a number of key factors which help to creatively regulate and stimulate motivational processes. These include capacious and precise 'linguistic' means of expressing mathematical regularities, the need for their justification and the development of a mathematical apparatus that enables a solution to any given problem of an applied nature. These factors can be described in hierarchical sequence [17, 18].

The initial link in this sequence is the practical need for solving specific pressing problems from the field of real human activity by the application of tailor made mathematical activity. In particular, it is expressed in the fact that any given practical task can stimulate the development of certain mathematical methods which in turn facilitate useful generalizations that would later extend those methods to a whole range of practical problems. This can itself become the starting point for relatively long-term mathematical research and open up the possibility of applying the developed mathematical theories to as yet unsolved practical problems. It can also help specify the prerequisites for the application of the developed mathematical tools to future stages of the development of science.

All these aspects should be taken into account when studying the school based mathematics curriculum. They offer a rich set of motivational opportunities. The particular motivational mechanisms which are applied will always initially be mindful of age range. In the initial stage of teaching mathematics, practical activity is basically a directly stimulating function in which the main role belongs to observation and experiment (real tasks for computing, measuring, plotting, constructing, etc.).

Motivational potential is then enhanced by demonstrating how a specific method for solving a specific problem can be used as the launching pad for the further development of the problem and for proving a theorem. The intention in so doing is to correlate this activity with the students' previous real life experience. Later, the motivational role of practice is realized through the recognition of the role of the studied mathematical content in the deployment of related courses and also through consideration of the history of the emergence and evolution of scientific concepts and methods. Further it develops understanding of the various elements of mathematical modeling of real states and processes which underpin the mastery of applied mathematical ideology. Textual tasks are at the core of the mechanism for such an implementation. All this continues at the most senior levels, but this is now combined with a growing awareness of ecological orientation – of the importance of the mathematic repertoire for general involvement in the world and in particular for the successful implementation of future professional activity. This function can be performed using real-world applications using a sufficiently serious mathematical tools.

The implementation of these mechanisms, reflecting the specifics of the implementation of the motivational role of practice in relation to mathematics, requires careful and precise correlation with each student's educational activity. The relevant material is presented in detail in our textbooks and articles.

References

[1] R. Courant, H. Robbins, What is Mathematics? An elementary approach to ideas and methods: London, Oxford University Press Incorporated, 1996.

[2] M. Kline, Mathematics. The Loss of Certainty: New York, Oxford University Press, 1980.

[3] M. Rodionov, The Formation and development of students motivation: Saransk, MGPI, 2012.

[4] M. Rodionov, Z Dedovets, " Practical Problems as Tools for the Development of Secondary School Students' Motivation to Learn Mathematics" in 19th International Conference on Computer Supported Education and Information Technology, New-Your, USA, 2017 https://waset.org/publications/10008026/

[5] G. Polya, Mathematics and Plausible Reasoning: Induction and analogy in mathematics: Princeton, New Jersey: Princeton University Press, 1954.

[6] L.C. Karpinsky, H. Y. Benedict, J. W. Calhoun, Unified Mathematics: Boston-New York-Chicago, D. C. HEATH & CO., 1918.

[7] T. Larkin, D. Budny, "Learning styles in the classroom: approaches to enhance student motivation and learning" in 6th International Conference on Information Technology Based Higher Education and Training, Santo Domingo, Venezuela, 2005. https://ieeexplore.ieee.org/document/1560310/

[8] M. B. Balk, V.A. Petrov, "About the mathematization of problems arising in practice" Mathematics at school, 3, 55-57, 1986.

[9] N. A. Tereshin, Applied orientation of school mathematics: M. Education, 1990.

[10] A. Ovezov, "Applying mathematical reasoning" Mathematics at school, 4, 45–52, 1998.

[11] A. D. Myshkis, M. M. Shamsutdinov, "Methods of applied mathematics" Mathematics at school, 2, 12-14.

[12] I. I. Blekhman, Mechanics and Applied Mathematics. Logic and applications of mathematics: M. Science, 1983.

[13] P. T. Apanasov, N. P. Apanasov, Collection of math problems with practical content: M. Education, 1987.

[14] H. Freudenthal, Mathematics as an Educational Task: Dordrecht-Holland, D. Reidel Publishing Company, 1973.

[15] W. Sawyer, Prelude to Mathematics: M. Education, 1972.

[16] I. Stewart, Concepts of Modern Mathematics: New York, Dover Publications, 1995.

[17] A. N. Tikhonov, D. P. Kostomarov, Stories about applied mathematics: M. Science, 1991.

[18] G. M. Wozniak, "The motivation problem in education" Mathematics at school, (2), 9-11, 1990.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 259-264 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Student Performance Evaluation Using Data Mining Techniques for Engineering Education

Veena Deshmukh[*,1], Srinivas Mangalwede[1], Dandina Hulikunta Rao[2]

[1]*Research Centre Gogte Institute of Technology Belgaum, India,*

[2]*Cambridge Institute of Technology Bengaluru, India*

ARTICLEINFO

Article history:
Received: 13 September, 2018
Accepted: 01 November, 2018
Online: 10 November, 2018

Keywords :
Mamdani Fuzzy Inference System
Scoring Rubrics Tool
Bloom's Taxonomy

ABSTRACT

In this research work, we are implementing a student performance evaluation model using Mamdani Fuzzy Inference System (FIS) and Neuro Fuzzy system and comparing the results with classical averaging method for Network Analysis (NA) course studied by third semester Electronics and Communication Engineering students. This work explains the designing of scoring rubrics using Bloom's levels as the criteria of assessment for NA course. Also at initial stages of learning how students' strengths and weaknesses can be identified using rubrics and develop critical thinking skills. The five inputs identify, understand, apply, analyze and design/create are five levels of learning as per Bloom's Taxonomy. Fuzzy rules are applied and the evaluated results are expressed in both crisp and linguistic variables and compared with classical aggregate scores.

1. Background and Motivation

Performance evaluation is always a challenge for a tutor after a learner undergoes a learning process or training program. Earlier the final grade for any course would mean the averaging of internal assessment scores and semester end examination scores. Later different activities like quiz, mini projects, seminars etc. were introduced in the course learning process to inculcate the higher order or critical thinking skills. In this process to the averaging method fails to highlight on the learners' critical thinking abilities. Also take a case of two students with same average marks but scores of both are like this, first student 65,75, 85 and second student 85,75,65. Here the average of the two students is same but first student is continuously improving the performance where as the second student is not. But a simple averaged grading method at the end of the course may not through light on all these issues of understanding, improvement or attainment of higher order thinking skills etc. [1]. The grades of evaluation of learners should in an ideal situation reflect their understanding of the course learnt, application of it in solving challenges of real life situations or solving similar issues by modifying the existing solutions or developing a new solution. Instead, continuous assessment of learner on daily/weekly basis is performed for given tasks and if a rubric with Bloom's criteria for evaluation on one side of the table and scores on the other side for predefined criteria is designed to evaluate the critical thinking skills for each of the tasks assigned then the rubrics scores clearly indicate the strengths and weakness of learner, levels of critical thinking skills attained, understanding of the course on daily/weekly basis [2,3]. The scores so obtained are then subjected to suitable soft computing evaluation technique where each level attained by learner is weighted. The final grade so obtained should give details of the rubrics scores highlighting these thinking abilities of a learner for the course learnt, the activities performed during the training period. Whenever the rubrics scores for daily/weekly (formative) assessments are discussed with the learner the learner understands the natural thinking abilities and they can be strengthened and if the learner is willing to work upon weak areas for improvement, then he/she can take guidance from the tutor and improve upon them. The rubrics scores for every task assigned can give early indications of strengths and weakness' to both learner and tutor. Such formative assessments provide scope for improvement in final grades and skills of learner and even tutor can plan new learner centric teaching methodologies.

The scoring rubrics [2,3,4] can be designed for each activity stating clearly the evaluation criteria/objectives to attain the standards set for the said activity. They can be used for formative and summative assessments. The grading is unbiased, more objective and the degree of attainment is easily understood by

[*]Veena Deshmukh, KLSGIT Belagavi Karnataka ,India, Contact No: +919916508826 & vbdeshmukh@git.edu

both learner and tutor. The rubrics can be modified during learning phase if required to improve standard of learning. The NA course demands identifying different circuit elements, understanding and applying Kirchhoff's Laws and finally calculating the various currents, voltages and designing equivalent networks. The different levels set to evaluate critical thinking skills are identify, understand, apply, calculate and create. The Bloom's Taxonomy levels are set as standards for evaluation.

Bloom's Taxonomy was designed in 1965 by Dr. Benjamin Bloom[5,6] to improve learning standards in higher education. As per this, three domains of learning are described. Namely Cognitive - Knowledge, Affective- Attitude and Psychomotor- Skills. The cognitive domain involves the development of intellectual skills and knowledge. This is further divided into six orderly categories. These are commonly referred as Bloom's levels and are in the increasing order of complexity and thinking abilities. Now it is referred globally to evaluate critical thinking skills, for defining course objectives, designing cognitive level course activities and tests, setting question papers at different levels etc.

In this work we have tried to overcome the drawbacks of averaging method by evaluating learner performance using scoring rubrics tool designed for Network Analysis course. The rubric is designed keeping the critical requirements of the said course like identifying different electrical elements of a network, analyzing the network for its loop and branch currents, node and branch voltages, related power calculations etc. Later, making use of this knowledge in creating or building new simple solutions for real life scenario. Also while designing rubrics tool to evaluate the learners' for critical thinking skills, Bloom's levels or higher order thinking skills are set as criteria/bench mark for evaluation. After designing the rubrics, it is discussed with the learners so that they understand the different criteria for evaluation. The score/grade card mentions the rubrics scores obtained indicating different levels attained by each learner for different tasks completed along with average/weighted average grade of performance (it may be CGPA –cumulative grade point average or percentage of scores). The soft computing techniques are investigated for suitability to evaluate the performance [7,8]. The Neuro-fuzzy (NF), Fuzzy inference system (FIS) are tested and performances are compared with classical averaging method for student performance evaluation. .

Professor L. A. Zadeh invented the concept of Fuzzy Logic in 1964. In 1974, Mamdani developed first fuzzy logic controller which is used in predicting results when data is imprecise, vague or some data is missing. Fuzzy controllers are widely used in forecasting weather, stock market, product market, health monitoring systems, aviation systems, temperature and pressure controllers in manufacturing industries etc. It is rule based and reliable.

2. Literature Review

A good number of researches have been reported since 1995 on student performance evaluation using soft computing techniques. Authors have proposed a fuzzy logic based model for performance evaluation of Network Analysis course [1,2] . The early work by Biswas used fuzzy sets for evaluation of students answer scripts by matching the answer scripts in 1995 [9]. But when large number of papers are to be evaluated this becomes tedious. Intelligent expert systems were implemented not using the complicated matching operation of answer scripts, instead [10] a more generalized method using degree of satisfaction and extended fuzzy grade sheets. A cricketer performance evaluation model was presented to predict international rank of a cricketer and also the effect of each input parameter like ranking of the teams being played, current ranking of the player, run rate etc are evaluated and rank of cricketer is predicted. Also the effect of each parameter on performance is discussed [11].The drawbacks of conventional method of evaluation performed in universities considering higher weightage to attendance is discussed in [12] also a new method using fuzzy logic considering student attendance as one of the parameters along with internal and external marks as input parameters is implemented. A personalized student evaluation method is presented in comparison to the back propagation and conventional statistical methods in [13]. Every student is unique and fuzzy systems can make decisions and evaluate student performance along with students learning progress Such a unique model is presented in [14]. In [15] the authors opine that the performance of the students depends upon their previous performance, medium of instruction and type of board affiliated to using WEKA tool and compare the results of various classifiers. There are several factors which affect the performance of students,[16] authors use the combination of Genetic Algorithm and Artificial Neural Networks to predict the performance and also to find the factors which influence the performance of students. In another paper authors predict performance considering gender, location of house, family income, medium of instruction along with previous semester grades, attendance [17]. The application of fuzzy inference system in predicting the traffic flows by considering traffic related parameters is discussed in [18]. Neuro fuzzy systems in which the different attributes are used to predict performance in crisp values. It also provides an alternate solution when data available is vague and imprecise and classify the students' performance into different categories [19,20]. Student performance evaluation and prediction can be done using Sugeno-type ANFIS architecture. The membership functions generated by ANFIS using the training data predict the student performance (21).Prediction and evaluation can also be done using different types of data mining algorithms like C5, CART, ANN,SVM (22).Ensemble is one of the data mining algorithm used to classify recorded heart sound in (23). Classification using KNN is studied using certain distance algorithms such as Cosine, Correlation, Euclidean and City blocks in (24).By considering students' previous performance, motivation level, family background future performance is predicted in to different classes as Very good, Good and Poor (25). The data mining techniques are used to obtain the hidden knowledge about the student performance (26).The performance of different data mining algorithms are compared by considering their 'No of True positive' values (27).

3. Methodology

In this work we design the rubrics for Network Analysis course as given in Table 2. for third semester Electronics and Communication Engineering students (ECE). This course deals with study of different electrical circuit analysis and simplification techniques like Mesh and Node analysis, Network Theorems for

DC and AC circuits etc. This requires the basic knowledge of electrical elements, Ohm's law and Kirchhoff's laws.

The steps involved are as follows:

- Designing of Rubric - the rubric is designed after consultation with course experts and students.
- Content delivery and Designing of question paper - the course content is delivered and question paper is designed as per Bloom;s taxonomy.
- Data collection - The test is conducted and scores of 41 students in rubric profile are collected.
- Evaluation using FIS and NF models - The data collected is investigated using the two models mentioned as shown in Figure 1.

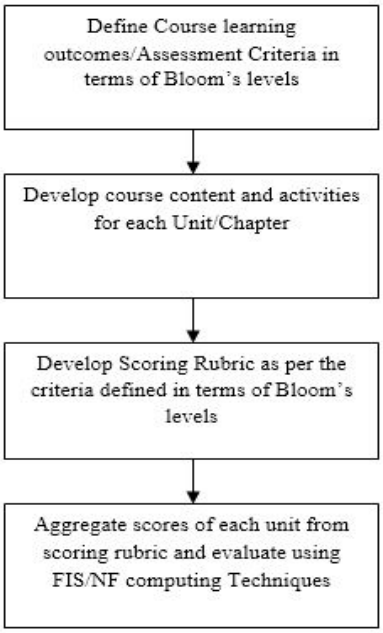


Figure 1. Designing of Scoring Rubric.

3.1. Fuzzy inference system (FIS)

The system which maps inputs to the outputs using fuzzy set theory is known as the Fuzzy Inference System. This can be either Mamdani or Sugeno. This system involves 3 steps, fuzzification of inputs using membership functions, formation of rule base using IF-THEN rules, defuzzification using output membership functions to get crisp values of outputs as shown in Figure 3. The membership functions can be chosen depending upon the requirement from the set of triangular, trapezoidal for linear variations and Gaussian, sigmoidal, zigmoidal etc.for nonlinear variations. We have used trapezoidal membership functions for both inputs and outputs as shown in Figure 2. Different weights are assigned to lower and higher order skills so that drawback of classical aggregation of scores is overcome. For example the scores above 8 are considered excellent for lower order skills like Identify whereas scores above 7 are considered excellent for hogher order skills like Apple and Create as shown in Table 2. The system can be made robust and flexible with the help of rules (13).

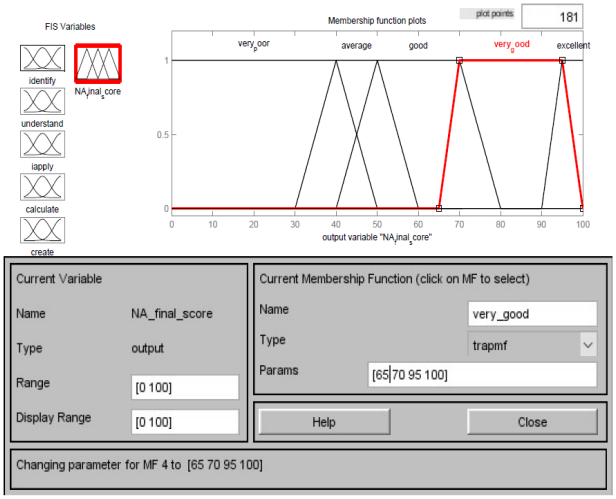


Figure 2. Membership functions for output.

Table 1. Input Variables

Input Variable	Evaluation Criteria
F1	Identify
F2	Understand
F3	Apply
F4	Calculate
F5	Create

3.2. Neuro-fuzzy (NF)

The hybrid of fuzzy inference system and neural networks is Neuro- Fuzzy (NF) system as shown in Figure 4. It combines the advantages of ANN and FIS in terms of ability to learn and think respectively making any system intelligent and think like human beings. This hybrid system basically is a neural network and is trained to generate IF-THEN fuzzy rules and membership functions of fuzzy systems. It is possible to incorporate common sense, intelligence and knowledge into the structure of neuro-fuzzy systems. The neuro-fuzzy system consists of 5 layers. The first one is the input and last one is the output layer and remaining three are the hidden layers responsible for generating membership functions, calculations and normalization (10).

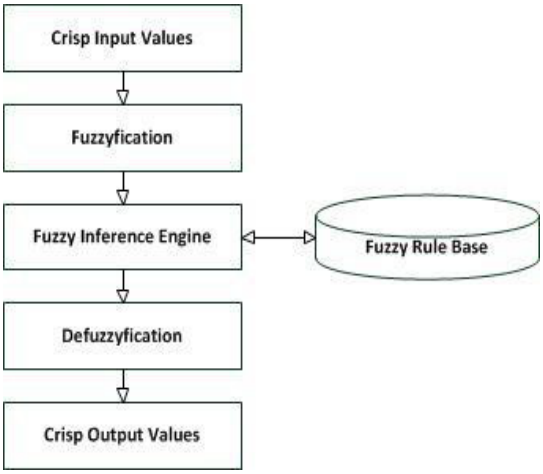


Figure 3. Basic flow of FIS

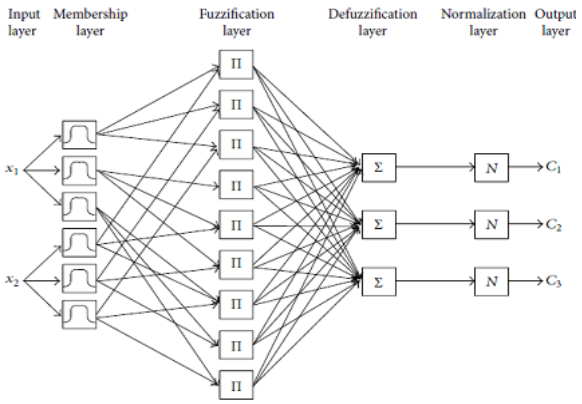


Figure 4: Basic bock diagram of NFZ

4. Results and Discussion

We have studied the rule based classifiers like Fuzzy (2), Neuro fuzzy systems which give 100% accuracy for classification. The crisp values for results are compared with classical averaging method. The initial time taken for these two systems is more if the data set is large for forming rule base in case of FIS or training the system in case of NF. Neuro fuzzy system and Sugeno fuzzy system are used with input variables as identity, understand, apply, calculate, create as shown in Figure 4. and Figure 5.The output variables indicate the performance of student's in linguistic variables and are expressed as Poor<5, Average>5, Good (6.0-6.9), Very good (7.0-8.90) and Excellent (>9.0) and are given in Table 4 Final results in Crisp values and linguistic variable are shown in Figure 6.

The comparison line graph of FIS, NF and classical averaging are shown in Figure 7. Comparison of different types of classifier results are as shown in Table 5.

Table 2. Final Scores in Linguistic Variables

Final score of NA(linguistic variable)	very poor	average	good	very good	excellent
Crisp value	<5.0	5.0-5.90	6.0-6.90	7.0-8.90	>9.0

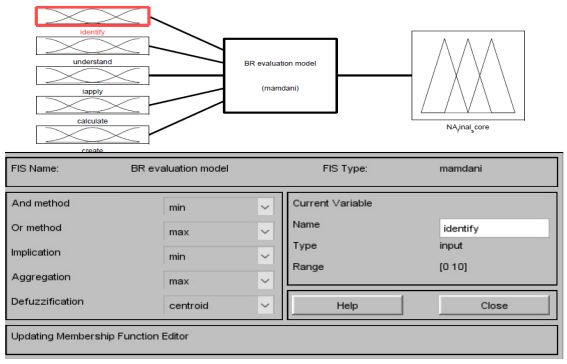


Figure 5. Fuzzy Inference System

5. Conclusions and future work

In this work we have compared the performance of Fuzzy Inference system and Neuro Fuzzy system with classical average scoring method. All these classifiers can be used to evaluate the student performance depending upon the accuracy of classification and time taken to classify. Depending upon the students attribute their performances are evaluated and classified into 5 classes as very poor, average, good, very good and excellent. We have compared the performance based on factors such as training time, accuracy of the classifier performance. Fuzzy inference and Neuro fuzzy systems give 100% classification accuracy and the results are comparable to classical averaging method. But if data is large then the training period is more and formation of rules, selection of membership functions lead to complexity. Depending upon the data size, accuracy required and training time constraint one can choose any of the above classifiers for performance evaluation. We would like to evaluate student performance using other soft computing techniques like Support Vector Machine,K- nearest Neighbour,Ensemble and Discriminant analysis in future.

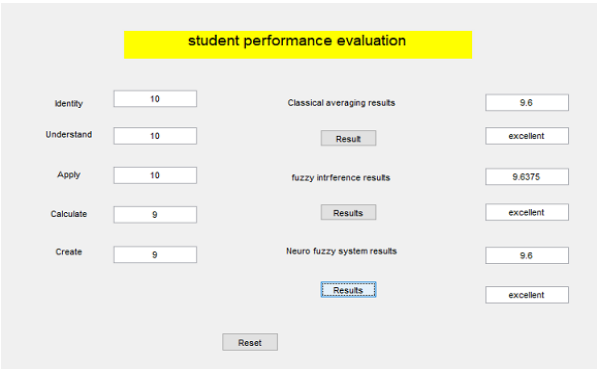


Figure 6. Result Comparison of different type of rule based classification

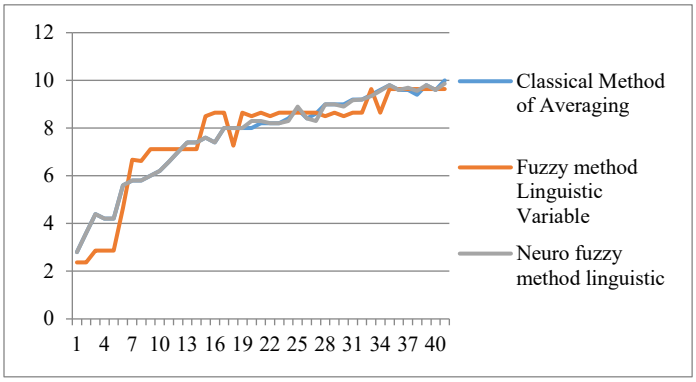


Figure 7. Observation Comparison Between the classical method, fuzzy and Neuro fuzzy system

Table 3. Different scores assigned for fuzzification of input variables

Input variables	unsatisfactory	satisfactory	good	very good	excellent
Identify	<5	5.0-5.9	6.0-6.9	7-7.9	>8.0
Understand	<5	5.0-5.9	6.0-6.49	6.5-7.49	>7.5
Apply	<5	5.0-5.49	5.5-5.9	6-6.9	>7.0
Calculate	<5	5.0-5.9	6.0-6.49	6.5-7.49	>7.5
Create	<5	5.0-5.9	6.0-6.49	6.5-6.9	>7.0

Table 4. Rubrics for Network Analysis Course

CLO or Criteria	Unsatisfactory(1-2)	Satisfactory(3-5)	Good(6-7)	Very Good(8-9)	Excellent(10)
Identify	Only Identifies and describes few circuit elements	Identifies and describes all the circuit elements	Identifies and describes the elements and demonstrates a limited understanding of the problem and solution	Identifies and describes the elements and demonstrates complete understanding of the problem and partial solution	Identifies and describes the elements and demonstrates complete understanding of the problem and solution
Understand/Select	Does not understand the problem including nodes and meshes	Does understand the problem including nodes and meshes	Selects and expresses nodes or meshes	Selects and expresses nodes or meshes in terms of equations	Selects and expresses nodes or meshes in terms of equations with complete understanding
Apply(KVL/KCL)	Does not apply mathematical/scientific principles/ laws	Applies laws with errors	Applies the mathematical/scientific principles/ laws with some errors and calculation mistakes	Applies the correct mathematical/scientific principles/ laws and all calculations are precise and not always appropriate	Applies the correct mathematical/scientific principles/ laws and all calculations are precise and appropriate
Calculate/ Compare	Fails to break down ideas/equations into simpler parts and analyze	Breaks down few ideas/equations into simpler parts	Breaks down ideas/equations into simpler parts, fails to analyze and compare/examine correctly	Breaks down ideas/equations into simpler parts analyzes and compares/examines sometimes correctly	Breaks down ideas/equations into simpler parts analyzes and compares/examines correctly
Summarize/Explain /Design	Does not acquire the knowledge in designing/developing electrical circuits	Acquires limited knowledge in only designing electrical circuits	Acquires limited knowledge in designing/developing electrical circuits	Acquires knowledge in only designing or developing electrical circuits	Acquires knowledge in designing and developing electrical circuits

Table 5.Student Data set.

Sl. No.	Identify	Understand/ Select	Apply	Calculate/ Compare	Create/ explain	Classical Method of Averaging	Fuzzy method results	Neuro fuzzy method results
1.	4	4	2	2	2	2.8	2.36944	2.80000
2.	4	4	3	3	4	3.6	2.36944	3.59999
3.	5	5	4	4	4	4.4	2.86609	4.39999
4.	5	5	3	4	3	4.2	2.86609	4.19997
5.	5	5	4	3	4	4.2	2.86609	4.19998
6.	7	7	5	4	4	5.6	4.64815	5.59999
7.	7	7	7	5	3	5.8	6.67544	5.79997
8.	8	8	5	6	2	5.8	6.62000	5.79999
9.	8	7	8	3	4	6.0	7.11905	5.99998
10.	6	7	6	6	6	6.2	7.11905	6.19995
11.	7	7	6	6	7	6.6	7.11905	6.59992
12.	7	7	7	7	7	7.0	7.11905	7.00001
13.	7	8	7	8	7	7.4	7.11905	7.40000
14.	8	7	8	7	7	7.4	7.11905	7.39999
15.	10	10	6	6	6	7.6	8.50000	7.59994
16.	10	10	6	6	5	7.4	8.64643	7.40000
17.	8	9	8	9	6	8.0	8.64643	7.99995
18.	8	8	8	8	8	8.0	7.26757	8.0000
19.	8	9	9	7	7	8.0	8.64643	8.00007
20.	10	9	10	7	8	8.0	8.50000	8.30003
21.	9	9	9	7	7	8.2	8.64643	8.29991
22.	10	10	7	7	7	8.2	8.50000	8.19828
23.	10	10	9	6	6	8.2	8.64643	8.20001
24.	9	9	9	8	7	8.4	8.64643	8.29991
25.	10	10	10	7	7	8.8	8.64643	8.89706
26.	10	10	10	6	6	8.4	8.64643	8.39996
27.	10	9	10	7	7	8.6	8.64643	8.30003
28.	9	10	9	7	8	9.0	8.50000	8.99999
29.	10	9	9	8	9	9.0	8.64643	8.99997
30.	10	10	10	7	8	9.0	8.50000	8.89706
31.	10	10	9	10	7	9.2	8.64643	9.17178
32.	10	9	9	10	8	9.2	8.64643	9.19997
33.	10	10	8	9	10	9.4	9.63750	9.34977
34.	10	10	10	8	10	9.6	8.64643	9.56348
35.	10	10	10	9	10	9.8	9.63750	9.80072

36.	10	10	9	10	9	9.6	9.63750	9.60004
37.	10	10	9	9	10	9.6	9.63750	9.69827
38.	10	10	9	9	9	9.4	9.63750	9.55849
39.	10	10	10	10	9	9.8	9.63750	9.80007
40.	10	10	9	10	9	9.6	9.63750	9.60004
41.	10	10	10	10	10	10	9.63750	9.86492

References

[1] V. B. Deshmukh, S. R. Mangalwede and Rao, D. H. "Student performance evaluation model based on scoring rubric tool for network analysis subject using fuzzy logic." In International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 2017, pp. 1-5. IEEE, 2017.

[2] V. B. Deshmukh, S. R. Mangalwede and Rao, D. H. "Student performance evaluation model based on Bloom's Taxonomy using fuzzy logic." In International Conference on Electrical, Electronics, Communication, Computer, Mechanical and Computing (EECCMC). IEEE, 2018.

[3] Srikant Jadcherla, " A short note on Revised Bloom's taxonomy" Seer Academi and Global University 2014.

[4] Kai Meng Tay, Chwen Jen Chen and Kim Khoon Lee, "Application Of Fuzzy Inference System To Criterion-Referenced Assessment With A Case Study", Proceedings of the 2nd International Conference of Teaching and Learning (ICTL 2009) INTI University College, Malaysia

[5] Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill,W.H., & Krathwohl, D.R. (1956). "Taxonomy of educational objectives: The classification of educational goals", Handbook 1: Cognitive domain.

[6] David R. Krathwohl. "Theory Into Practice", Volume 41, Autumn 2002 The Ohio State University.

[7] Brenda Sugrue, October, 2002 " Problems with Bloom's Taxonomy".

[8] Zadeh, L.A. 1965, "Fuzzy sets and systems", In J. Fox, editor, System Theory. New York: Polytechnic Press, pp. 29-39.

[9] Biswas, "An application of fuzzy sets in students' evaluation" 1995, Fuzzy sets and systems pp.187-194.

[10] Shyi-Ming Chen, Chia-Hoang Lee, "New methods for students' evaluation using fuzzy sets", Fuzzy sets and Systems 1999, pp.209-218.

[11] Singh, G., Bhatia, N., and Singh, S. "Fuzzy Logic Based Cricket Player Performance Evaluator". IJCA Special Issue on "Artificial Intelligence Techniques - Novel Approaches & Practical Applications" 2011, pp. 11-16.

[12] Meenakshi N., Pankaj N., "Application of Fuzzy Logic for Evaluation of Academic Performance of Students of Computer Application Course", IJRASET 2015. Volume 3 Issue X, October 2015 ISSN: 2321-9653 .

[13] Nidhi Arora, Jatinder Kumar R. Saini. "A Fuzzy Probabilistic Neural Network for Student's Academic Performance Prediction", International Journal of Innovative Research in Science, Engineering and Technology, 2013,Vol. 2. 4425-4432.

[14] Z.Jeremic,J.Jovonovic,D.Gasevic, "Sudent modeling and assessment in intelligent tutoring software paterns", Expert systems with Applications, Elsevier, 2012 pp.210-212 .

[15] R. Sumitha , E.S. Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques", International Journal of Scientific Engineering and Applied Sciences , Volume-2, Issue-6,June 2016 ISSN: 2395-3470.

[16] Omar Augusto Echegaray-Calderon, Dennis Barrios-Aranibar, "Optimal selection of factors using Genetic Algorithms and Neural Networks for the prediction of students' academic performance", 978-1-4673-8418-6/15/$31.00 ©2015 IEEE.

[17] Gurmit Kaur,Williumjit Singh, " Prediction of Student Performance Using WEKA Tool",An International Journal of Engineering Sciences,2016,Vol 17, ISSN:2330-0332(online).

[18] Deshpande, Minal, and Preeti R. Bajaj. "Short term traffic flow prediction based on neuro-fuzzy hybrid sytem." In ICT in Business Industry & Government (ICTBIG), International Conference on, pp. 1-3. IEEE, 2016.

[19] Isa, K., Mohamad, S. and Tukiran, Z., 2008, August. Development of INPLANS: An Analysis on Students' Performance using Neuro-Fuzzy. In Information Technology,ITSim 2008. International Symposium on (Vol. 3, pp. 1-7). IEEE. 2008

[20] Sevarac, Zoran. "Neuro fuzzy reasoner for student modeling." In Advanced Learning Technologies, 2006. Sixth International Conference on, pp. 740-744. IEEE, 2006.

[21] Taylan, Osman, and Bahattin Karagözoğlu. "An adaptive neuro-fuzzy model for prediction of student's academic performance." Computers & Industrial Engineering 57, no. 3 (2009): 732-741

[22] Agaoglu, Mustafa. "Predicting instructor performance using data mining techniques in higher education." IEEE Access 4 (2016): 2379-2387.

[23] Homsi, Masun Nabhan, Natasha Medina, Miguel Hernandez, Natacha Quintero, Gilberto Perpiñan, Andrea Quintana, and Philip Warrick. "Automatic heart sound recording classification using a nested set of ensemble algorithms." In Computing in Cardiology Conference (CinC), 2016, pp. 817-820. IEEE, 2016.

[24] Kataria, Aman, and M. D. Singh. "A review of data classification using k-nearest neighbour algorithm." International Journal of Emerging Technology and Advanced Engineering 3, no. 6 (2013): 354-360

[25] Do, Quang Hung, and Jeng-Fung Chen. "A neuro-fuzzy approach in the classification of students' academic performance." Computational intelligence and neuroscience2013 (2013): 6.

[26] Amra, Ihsan A. Abu, and Ashraf YA Maghari. "Students performance prediction using KNN and Naïve Bayesian." In Information Technology (ICIT), 2017 8th International Conference on, pp. 909-913. IEEE, 2017.

[27] Troussas, Christos, Maria Virvou, and Spyridon Mesaretzidis. "Comparative analysis of algorithms for student characteristics classification using a methodological framework." In Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on, pp. 1-5. IEEE, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 260-269 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Topology, Discontinuities and Dimension Effects on CMOS Rotary Traveling Wave Oscillators

Rogelio Manuel Higuera Gonzalez [*,1], Mónico Linares Aranda[2], Luis Hernández Martínez[2], Carlos Sanabria Díaz[2]

[1]*Electronics Department, Tecnológico de Estudios Superiores de Ixtapaluca (TESI), 56580 México*

[2]*Electronics Department, Institute of Astrophysics, Optics and Electronics (INAOE), 72840 México*

ARTICLEINFO

Article history:
Received: 30 June, 2018
Accepted: 31 July, 2018
Online: 12 August, 2018

Keywords:
RTWO
Oscillator
Transmission lines

ABSTRACT

In this article, a new method to extract the RLC parameters from the structure of a Rotating Traveling Wave Oscillator (RTWO) is described. The use of a 3D simulator (EMPRO) allow us to develop a precise analysis of the impact of the topology, the geometrical discontinuities, and the dimensions of the transmission lines, on the performance parameters of the RTWO, i.e. frequency of oscillation and power consumption. The results show that a right angle discontinuity compensation can reduce the power consumption up to 3.34% and, for RTWOs with a frequency range between 11.4 GHz and 11.5 GHZ, narrow transmission lines (as the technology is able) can reduce the power consumption up to 98%. Additionally, it was observed that the topology used affects the performance parameters, as there are more folds and overlaps between metal levels in the RTWO structure, and the frequency increases by 2.9% while the power consumption increases by 10.52%.

1 Introduction

Nowadays, the high and very high-frequency ranges (millimeter range: 10-66Hz, and THz range: 0.1THz-10THz) are an object of study due to their promising applications in the military, commercial, communications and medical fields; which include spectroscopy, remote environmental sensing, high-resolution biological imaging, mobile telecommunications, and broadband satellite communications, among others [1, 2, 3]. On the other hand, in the past, only critical circuits, essentially high-precision clock systems, operated at high frequency.

However, this is no longer true at the present high-speed digital design. Many of the signal processors and their interconnections operate and transmit signals at frequencies in the MHz range and even in the GHz range. This is the case of high-performance information systems, which after the current demand of developing faster and faster electronic devices able to perform a wide number of functions (computers, telephones, tablets, and so on), generate the need to increase the operating frequency of the microprocessors to the GHz range with a minimum power consumption. Consequently, the conventional systems of clock generation and/or distribution, commonly used in the current ICs (H-tree, X-tree, mesh, etc.), become obsolete.

The common element used in the previously mentioned applications is the oscillator: a periodic signal generator circuit that operates at the highest frequency in any synchronous system. In communications systems, the clock element is the Voltage Controlled Oscillator (VCO), usually included in the Phase Tie Loop (PTL) and in microprocessor systems. Currently, there is a new resonant oscillator type based on transmission lines. To synchronize signals on CMOS chips, different high-speed resonant oscillator structures have been created: distributed differential oscillators [4], fixed wave oscillators [5], and Rotary Traveling Wave Oscillators (RTWO) [6], among others. From these technologies, the RTWO stands out for its multiple advantages such as high oscillation frequencies, low power consumption, multiphase signals, low jitter, low skew and simple design [7].

The RTWO is an innovative approach which takes

[*]Rogelio Manuel Higuera Gonzalez, Carretera Coatepec s/n Km. 7, San Juan, 56580 Ixtapaluca, México, (+32) 555-477-3941 & m.sc.hgrm@ieee.org

https://dx.doi.org/10.25046/aj030426

advantage of the distributed Inductive and Capacitive (LC) nature of the interconnection lines, present at the CMOS processes used for the manufacture of integrated electronic systems. RTWOs are used for the generation of multi-GHz signals in communication systems, and the generation/distribution of clock signals in high-performance microprocessors [6].

The RTWO generates the oscillations through the LC elements associated with the transmission lines which form the resonant structure. Hence, the operation and performance of an RTWO depend on the physical properties of the materials used, the dimensions of the transmission lines (width W, length l, and thickness T), the discontinuities of the line (curves, crossings, vias, joints, etc.), and the topological structure used (circular, square, hexagonal, etc.).

Several works have been published concerning the design of the RTWOs. In [11], a rectangular topology was implemented and analytical approaches were used to obtain the RLC parameters. [19, 20] implemented different RTWOs in a 180 nm technology with a frequency range from 2GHz to 12GHz, using a rectangular topology and a 2D electromagnetic simulator to extract the RLC parameters. In [17], RTWOs were designed in a 65nm technology achieving a frequency range from 110 GHz to 152GHz. That work implemented an octagonal topology while the extraction of the RLC parameters was done by means of a 2D electromagnetic simulator. [23] used a 180nm technology with a frequency range from 3GHz to 11GHz, and although the extraction of the RLC parameters was done by a 3D electromagnetic simulator, they segmented the ring to perform the simulation, so they did not take into account many of the effects obtained when the complete ring is simulated.

Most of the mentioned previous works used analytical approximations or 2D simulators for the extraction of the RLC parameters. Since the analytical approximations are not exact and the 2D simulators do not take into account some electromagnetic effects on the ground plane, such as skin effect and proximity effect, the designed RTWO could not oscillate at the desired frequency or, in the worst case, could not oscillate.

In this work, a new design method was developed. A 3D electromagnetic simulator was used to extract the RLC parameters of the RTWO structure, which allows us to perform an accurate analysis of how the frequency of oscillation and power consumption are affected by the dimensions of the RTWO (length, width and gap between the transmission lines), right-angle discontinuities, and the different topologies used in RTWO design. In this way, it is possible to simultaneously analyze the effect of the geometric discontinuities and the topology on the operating parameters of the RTWOs.

2 The RTWO

The Rotatory Traveling Wave Oscillator (RTWO) is a closed loop implemented with a differential transmission line (displaying LC characteristics) with a connected crossing from the inner conductor to the outer

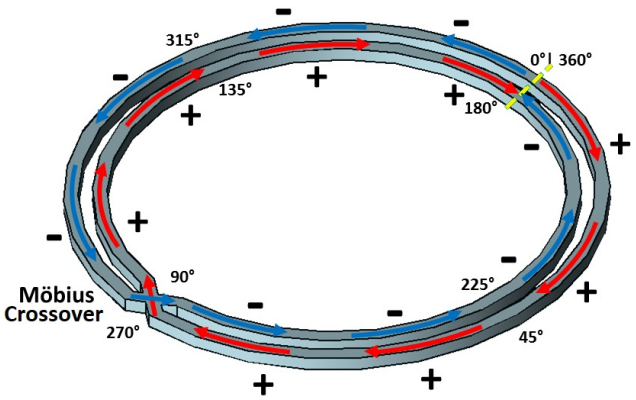


Figure 1: Typical circuit topology of a RTWO.

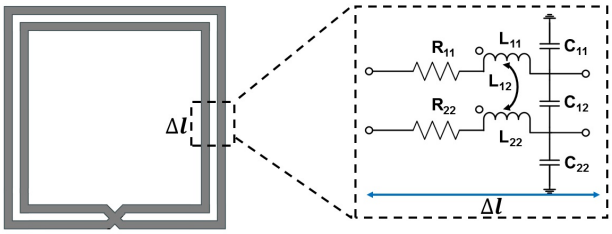


Figure 2: Equivalent electrical circuit of an infinitesimal section of an RTWO.

conductor (Möbius crossing), which causes a reversal in the signal (Figure 1). Oscillation begins with any noise event. If there were no losses, a wave would travel in the ring indefinitely, providing a complete clock cycle every two rotations to the ring (Möbius effect).

To model the electrical behavior of the RTWO resonant circuit, an equivalent electrical circuit of distributed elements (RLC) is used. The effects of the transmission line are modeled by resistance (R_{11} and R_{22}), self-capacitance (C_{11} and C_{22}) and self-inductance (L_{11} and L_{22}), while the magnetic and electrical coupling are modeled by mutual inductance (L_{12}), and mutual capacitance (C_{12}), respectively. The equivalent electrical circuit is shown in Figure 2.

Since there are losses in the conductor due to discontinuities or the dielectric, the signal will be attenuated while it travels through the resonant circuit. Therefore, an amplification stage is required to compensate the losses of the resonant circuit. For the amplification stage, a topology consisting of two inversers antiparallely connected (CCIP) were used. Additionally, the CCIPs block the phase, which causes the odd propagation mode, i.e. a signal in one line will be offset by 180° from the signal in the other line. The compensation stage configuration is shown in Figure 3. The CCIP stages will generate negative resistance

(transconductance), which compensates the resistive losses of the resonant circuit. These compensation stages are evenly redistributed throughout the ring.

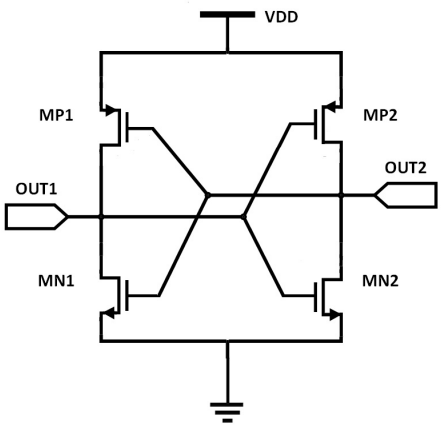


Figure 3: RTWO compensation stage.

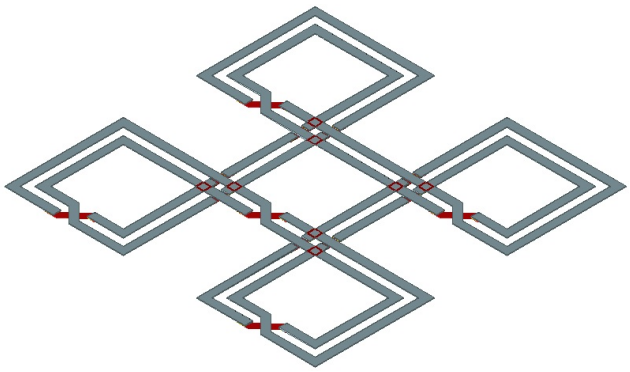


Figure 4: Rotary Oscillator Array (ROA).

2.1 Rotary Oscillator Array (ROA)

Multiple RTWOs can be connected together to form an array of rotating oscillators that distribute a signal to the entire chip. Figure 4 shows a basic structure of a ROA constituted by five RTWO rings connected in a checkerboard topology.

The inner rings and outer rings of the adjacent RTWOs are physically connected for synchronization purposes [14]. The advantage of implementing a ROA instead of a single RTWO is that a ROA provides more interconnection points, which eliminates the need of using too long interconnections between the synchronous element and the oscillator. The decrease in interconnection length provides the reduction in signal delay caused by the effects of high-frequency interconnections.

2.2 RTWO design parameters

Two of the main design parameters of a RTWO are power consumption and frequency of oscillation, which will be described in the following sections.

2.2.1 Power consumption

The power consumption of the RTWO is the sum of the contribution of the resonant circuit and the compensation stages, where the power dissipated by the resonant circuit P_{lt} can be calculated by Equation 1:

$$P_{lt} = \frac{{V_{DD}}^2}{{Z_c}^2} \tag{1}$$

The power consumption of the compensation stage P_{CCIP} is divided into three components: the static power consumption, which is negligible; the dynamic power, which is recirculated and converted into energy for the resonant circuit; and the short circuit power, which cannot be ignored and is consumed when both NMOS and PMOS transistors are switched on [15]. P_{CCIP} is expressed by Equation 2 as:

$$P_{CCIP} = \frac{I_s V_{DD}(t_r + t_f)}{T} \tag{2}$$

Where I_s is the saturation current, t_r is the signal rise time, t_f is the signal fall time and T is the signal period.

2.2.2 Frequency of oscillation

The frequency of oscillation f_{osc} is a factor of the total capacitance C_t and the total inductance L_t, as is expressed in Equation 3.

$$f_{osc} = \frac{1}{2\sqrt{L_t C_t}} \tag{3}$$

L_t is just related with the differential transmission line, while C_t depends on Equation 4. C_{lt} is the capacitance of the differential transmission line, C_{comp} is the capacitance of the compensation stage and C_{load} is the load capacitance.

$$C_t = C_{lt} + C_{comp} + C_{load} \tag{4}$$

As is seen from Equations 3 and 4, f_{osc} will strongly depend on the geometry of the RTWO and the compensation stage.

2.3 RTWO structure and manufacture

The RTWO of the present work was built using a 180 nm mixed-mode UMC manufacturing technology, which uses a single level of polysilicon and 6 levels of metal (1P6M), P-type substrate, and RFCMOS. The structure was made of aluminum at level 6, with the Möbius crossing at level 5. Figure 5 shows a schematic of the technology used. The main physical and electrical parameters of the RTWO are shown in Table 1.

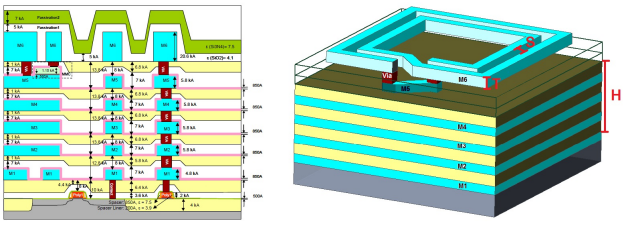


Figure 5: (a) Cross-section view of the UMC 180 nm RFCMOS manufacturing technology, (b) 3D structure of the RTWO.

Table 1: UMC 180 nm CMOS manufacturing process parameters, and RTWO geometrical features.

Parameter	Symbol	Value
Conductivity (S/m)	σ	37×10^6
Material permitivity	ε_r	3.8
Vacuum permitivity (F/m)	ε_0	8.854×10^{12}
Vacuum permeability (H/m)	μ_0	$4\pi \times 10^{-7}$
Line thickness (μm)	T	2.06
Oxide thickness (μm)	H	6.52
Line length (μm)	l	1000

2.4 Geometrical discontinuities

The abrupt changes in the conductor geometry of the transmission line cause geometrical discontinuities. Consequently, the electric and magnetic field distributions are modified near the discontinuity when high-frequency signals are transmitted through them.

In RTWOs design, right-angled corners origin a build-up of load particularly on the outside of the corner, which causes capacitance overload. Similarly, the interruption of the current flow produces inductance overload [8, 9]. Since the excess of capacitance and inductance change the frequency of oscillation and the power consumption of the RTWO, it is necessary to use a compensation technique for this type of geometrical discontinuity.

There are different useful techniques to compensate the excess of capacitance and inductance caused by 90° corners. These techniques remove some of the metal from the corners. Figure 6 shows the full-wave results from a geometrical discontinuity simulation. The current density is best redistributed by a 45° chamfer, 1 side [16]. Therefore, the less load build-up at the corners decreases the capacitance, compared to a right angle corner.

2.5 Topologies

In RTWO design, several topologies of different shapes are known, such as circular, hexagonal, octagonal, and square with different number of crossings. These topologies have been applied in [10, 17, 18, 21]. However, there is no analysis joining the effect of the diffrent discontinuities on the answer of an RTWO. In RTWO design, the dimensions of the oscillator, the design of the compensation stage, the geometric discontinuities, and the topology used for the oscillator, have a high impact on the inductance and capacitance of the differential transmission lines from which the oscillator is built. As consequence, they affect the oscillation frequency and the power consumption of the RTWO.

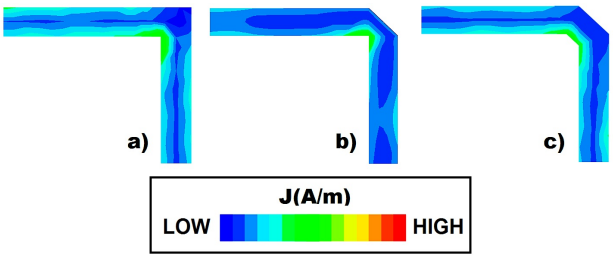


Figure 6: Current distribution on transmission line using mitered techniques: (a) Right angle bend (no mitered), (b) 45° chamfer - 1 side, (c) 45° bend - 2 sides.

3 Methodology

In order to include the effects of the geometrical discontinuities, in [19] and [23] the transmission lines of the rotating ring were divided into uniform segments. However, the union of the different segments presents differences in contrast with the superposition of the effects by segments. Both technique differ from the simulation of a complete ring [22]. Thus, to minimize the discrepancies and to consider the greatest number of electromagnetic effects in the structure of a RTWO, the methodology used in this work includes the simulation of the full ring.

Different software tools were used to simulate the RTWO due to the fact that the 3D electromagnetic simulator used (EMPRO) only extracts the S-parameters from the full-wave simulation. Then, it requires the use of an additional software to calculate the parameters of the RTWO through the S-parameters (MATLAB in this work), and another one to simulate the electrical circuit of the RTWO (HSPICE). Therefore, in this work the first step was to use EMPRO to develop the 3D geometrical structure of the RTWO. Then, a full-wave simulation was performed by means of the Finite Element Method (FEM). The resulting S-parameters from the simulation were exported to MATLAB. With MATLAB, the RLC parameters, the transistor size of the compensation stage, the power consumption and the oscillation frequency by analytical approaches were calculated. Finally, HSPICE was used to implement the RTWO considering the RLC parameters of the resonant circuit and the compensation stage, simulating the circuit of the distributed elements of the resonant circuit and the compensation stages of the RTWO.

At first, a working frequency range from 11 to 12 GHz was defined. In the simulations of [22], it was determined that with a length of 1000μm, the frequency

of oscillation range required was obtained. Therefore, different simulations of the RTWO were carried out varying the width W, the separation between the transmission lines S, and the topology of the RTWO, to observe the impact of these factors over the oscillation frequency and the power consumption. In the simulations developed with HSPICE, eight uniformly distributed amplification stages were used.

4 Effect of W and S on f_{osc} and P_t

All the electrical properties of the interconnections are due to the arrangement of the conductors and dielectrics and the way they interact with the electrical and magnetic fields of the signals. The total resistance of the interconnections depends on the signal strip and the return plane; both resistors depend on the geometry of the conductors and the current distribution in them. The inductance is also composed of two terms, the first related to the inductance inside the conductors and the second term, called external inductance, produced in the proximities of the conductors. Both inductances, as well as the resistance, depend on the current distribution in the conductors. On the other hand, the value of the capacitance depends on the distribution of the charge on the surface of the conductors.

As it was seen, inductance, resistance and capacitance in the interconnections, depend strongly on their geometry. Since the resonator circuit of the RTWO is built by the interconnections of the integrated circuits, and the frequency of oscillation depends on the capacitive and inductive effects of the interconnections, the impact of the width of the interconnections on the frequency of oscillation and the power consumption of the RTWO must be analyzed. Therefore in this work, several simulations were performed considering W in the range of 5 μm to 25 μm. Moreover, the value of S was set at the minimum value of W, in this case, 5 μm.

Figure 7 shows the variation of the total inductance L_t and the total capacitance C_t for the different values of W. As W increases, the cross-sectional area over which the current is distributed increases, so the total number of magnetic field line loops decreases and, therefore, L_t decreases. When increasing the size by W, the transconductance g_m required to compensate the losses must increase, so larger transistors are required and, as consecuence, C_{comp} increases. In the same way, considering the analogy of a parallel plate capacitor, when increasing the width of the line, the capacitance C_{lt} increases. The increment of both capacitances (C_{comp} and C_{lt}) produce the increment of the total capacitance.

Figure 8 shows the variation of the characteristic impedance Z_c and the resistance R when varying W. In this case, it can be seen that both parameters decrease with the increase of W because there is a larger cross-section where the current flows through the transmission line. Figure 9 shows how the frequency of oscillation and power consumption varies by increasing W. The low dependency of f_{osc} to W is due to the fact that the increase of C_t is compensated by the decrease of L_t (Figure 7).

Setting the values of l=1000 μm and W=5 μm, another parameter that can be varied is the separation between the transmission lines S. In this case, values from 5 μm to 15 μm were taken for S. Figure 10 shows the variation of L_t and C_t depending on the different values of S.

Due to the separation of the transmission lines, a smaller number of magnetic field lines are annulled, which causes the increase of the mutual inductance and therefore the increase of L_t. On the other hand, as the transmission lines are separated, fewer losses are required to be compensated for, so the size of the transistors decreases and C_t decreases. Figure 11 shows the variation of R and Z_c depending on the different values of S. It can be seen from the graph that both R and Z_c have little dependence on the separation of the transmission lines.

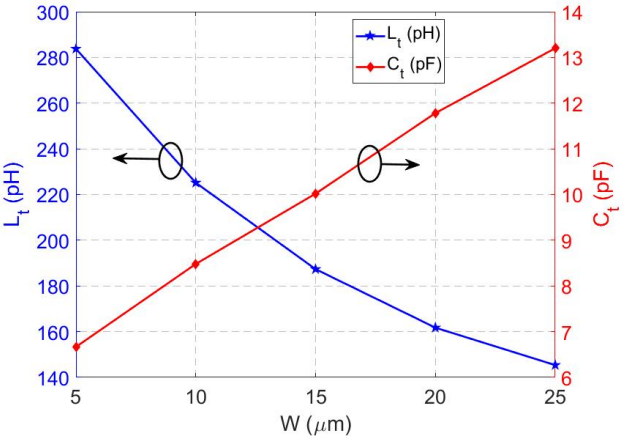


Figure 7: Variation of L_t and C_t when increasing W.

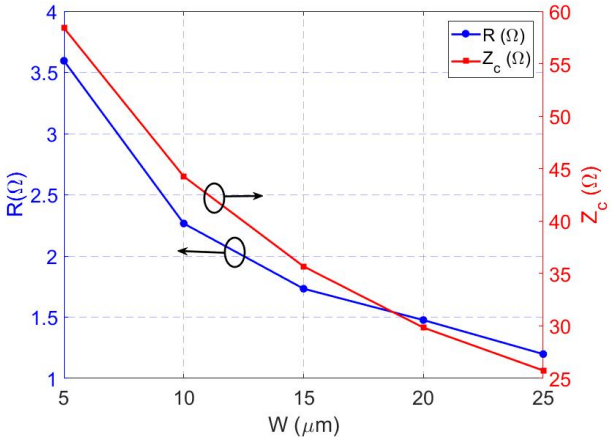


Figure 8: Variation of R and Z_c when increasing W.

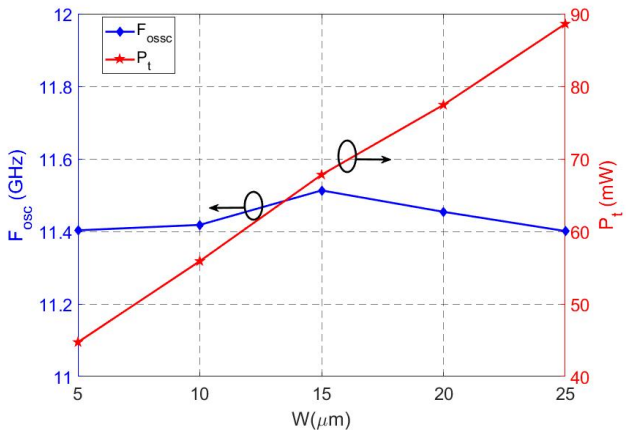


Figure 9: Variation of f_{osc} and P_t when increasing W.

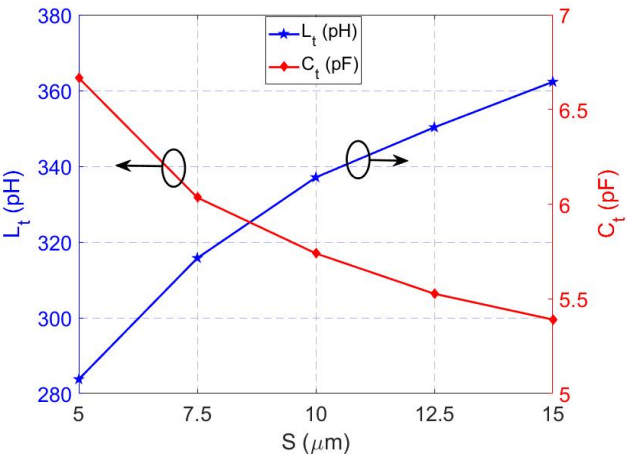


Figure 10: Variation of L_t and C_t when increasing S.

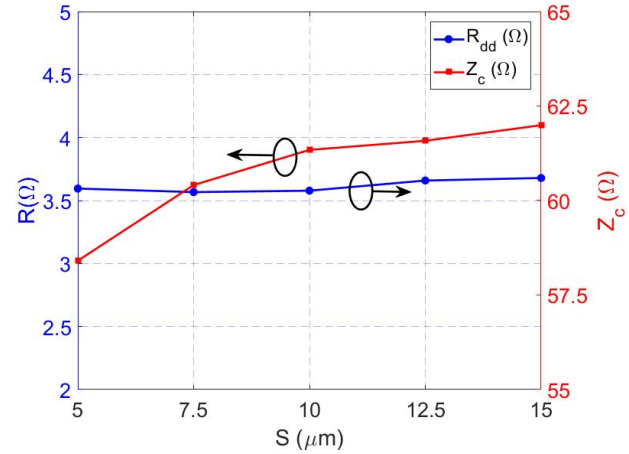


Figure 11: Variation of R and Z_c when increasing S.

Finally, Figure 12 shows that f_{osc} remains approximately constant. On the other hand, the power consumption P_t decreases when separating the transmission lines, due to the decrease of the power consumption of the compensation stage.

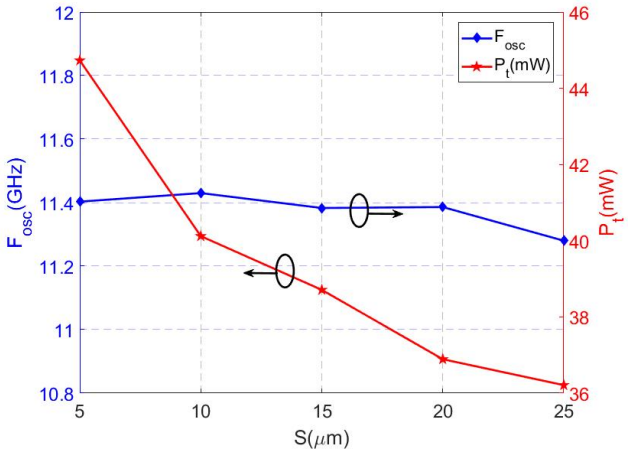


Figure 12: Variation of f_{osc} and P_t when increasing S.

5 Discontinuities compensation effect in squared topology

Prior to the simulation of the different RTWO topologies, the squared RTWO design was optimized by applying two mitered techniques: 45° miter - 1 side, and 45° miter - 2 sides (Figure 6). Three layouts were developed using the squared topology with a Möbius crossing, by means of the Mentor Graphics Design Editor (Figure 6a). In Layout 1 no mitered was implemented, but in Layout 2 the 45° mitered corner - 1 side was implemented, and in Layout 3 the 45° mitered corner - 2 sides. For the hexagonal and octagonal layouts, any mitered technique was applied. The performance parameters of the oscillators applying the different mitered techniques are shown in Table 2. It is worth to mention that Layout 4 corresponds to the squared topology of the RTWO built directly with the EMPro electromagnetic simulator, and not with Mentor Graphics, but for ease, it is also included.

Table 2: RTWO performance parameters using different mitered techniques.

Layout	F_{osc}**(GHz)**	P_t **(mW)**
1	11.66	41.18
2	11.82	39.28
3	12.07	37.74
4	11.43	40.12

From Table 2 is inferred that the 45° mitered on both sides of the line (Layout 3), is the most effective mitered technique because it allows to the RTWO having the highest oscillation frequency, in agreement with [16], due to its less capacitance and total inductance variation. In the same way, when comparing values of extracted capacitance, there is a variation of 5% in the total capacitance where EMPro is used directly (Layout 4) and when the RTWO layout is done in Mentor Graphics (Layout 3). This variation is due to the fact that in the layout from Mentor Graphics several vias (small contacts between the different metals of the technology) were used to connect the 5th and

6th levels of metal, while in Layout 4 only one single contact is used, which is unsuitable for real integrated circuits.

6 Topology effect

Figure 13 shows the total variation of the resistance, inductance, and capacitance for the different topologies of the RTWO analyzed. As is shown, the total capacitance increases since there are more crossings in the metal 5 (see topologies in Figures 14d to 14f). This effect is due to the increase in capacitance coming from the crossing. Figure 14(b) shows how the crossing in metal 5 (topologies 4, 5 and 6) increase resistance due to the smaller thickness of the metal 5 in contrast with metal 6. Therefore, the cross-section through which the current flows is lower causes an increase in the resistance.

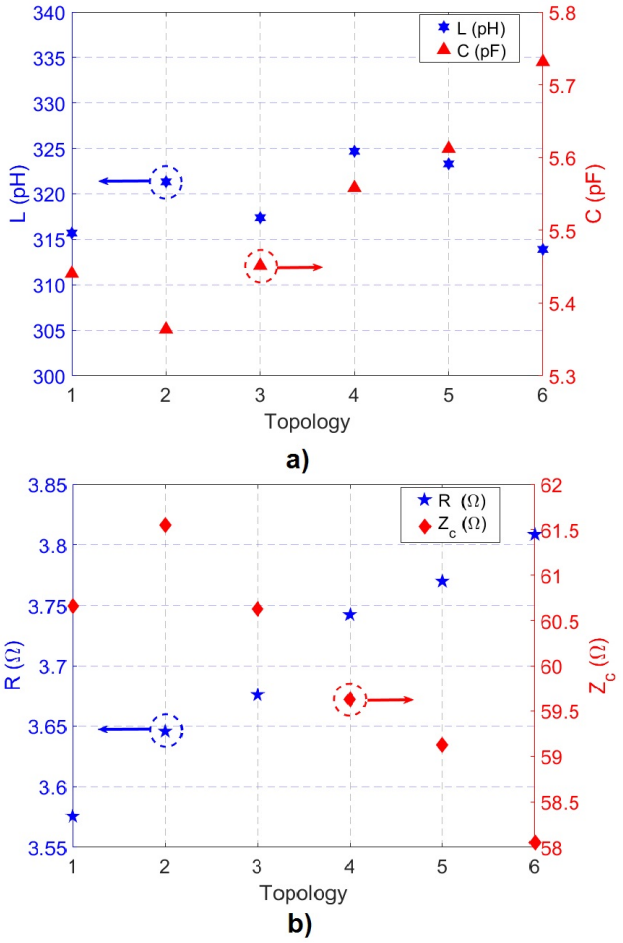


Figure 13: Electrical parameters of the different topologies: Resistance (R), Inductance (L), Impedance (Zc) and Capacitance (C).

Table 3 shows the performance parameters of the RTWO using different topologies. In this table, two crossing types are identified: one where metal levels 5 and 6 intersect diagonally (Möbius crossing), and the other one where metal levels 5 and 6 intersect orthogonally (simple crossing). It is observed that the oscillation frequency between oscillators having only one metal crossing (topologies 1, 2 and 3) is approximately the same; whereas if the number of crossings in metal 5 increases, the oscillation frequency decreases due to the increase of the inductance and the capacitance in the crossings. On the other hand, the power consumption increases as the number of metal 5 crossings increase (topologies 4, 5, and 6), due to the increase of the resistance and the reduction of the characteristic impedance.

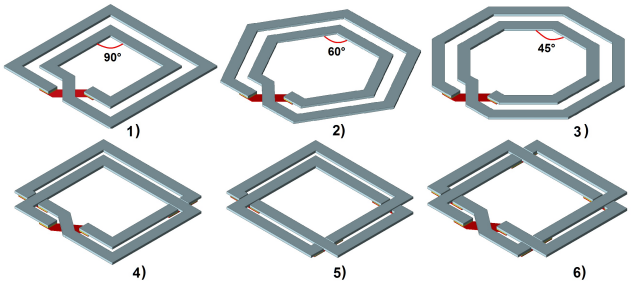


Figure 14: RTWO topologies analyzed: 1) Square with 1 Möbius crossing, 2) Hexagonal with 1 Möbius crossing, 3) Octagonal with 1 Möbius crossing, 4) Square with 2 crossings and 1 Möbius crossing, 5) Square with 3 crossings, 6) Square with 4 crossings and 1 Möbius crossing.

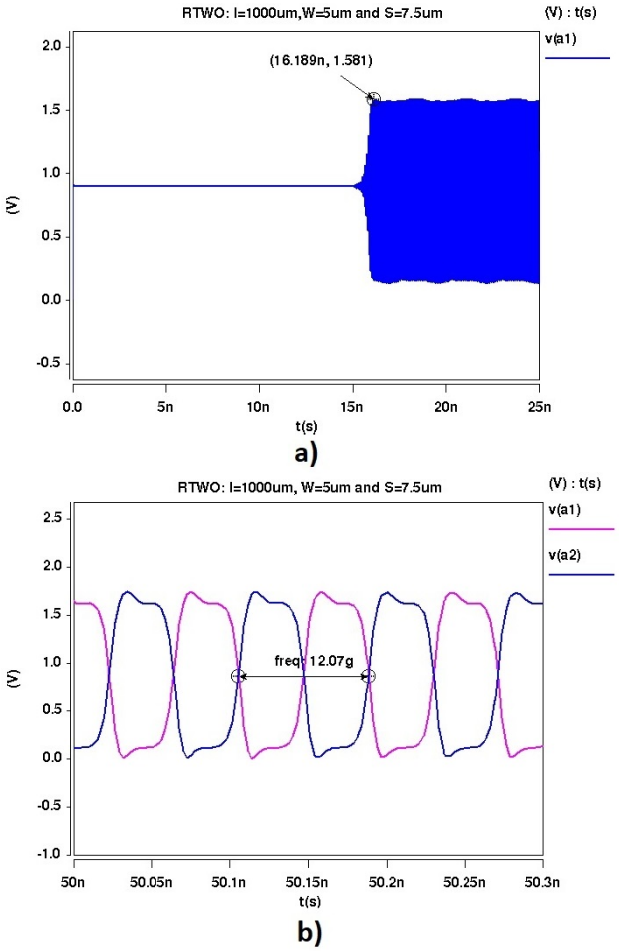


Figure 15: Waveforms of the RTWO oscillator: a) Beginning of the oscillation, b) oscillation of stable complementary nodes.

Table 3: RTWO performance parameters using diferent topologies.

Topology	F_{osc}(GHz)	P_t (mW)
1 (Squared-1 Möbius)	12.051	37.44
2 (Hexagonal-1 Möbius)	11.972	39.00
3 (Octagonal-1 Möbius)	11.937	39.67
4 (Squared-2 crossings-1 Möbius)	11.706	40.40
5 (Squared-3 crossings)	11.741	39.85
6 (Squared-4 crossings-1 Möbius)	11.701	42.18

Figure 15 shows the waveforms of the RTWO. In Figure 15a, is observed that, after 16.189 ns, the RTWO starts to oscillate in a stable form to a frequency of 12.07 GHz with an amplitude of 1.581 V. Figure 15b shows how the waveform of a node is offset by 180° from its complementary node. To observe the band nature of RTWO, the spectrum was observed over a span of 100GHz. Figure 16 shows the results of this simulation.

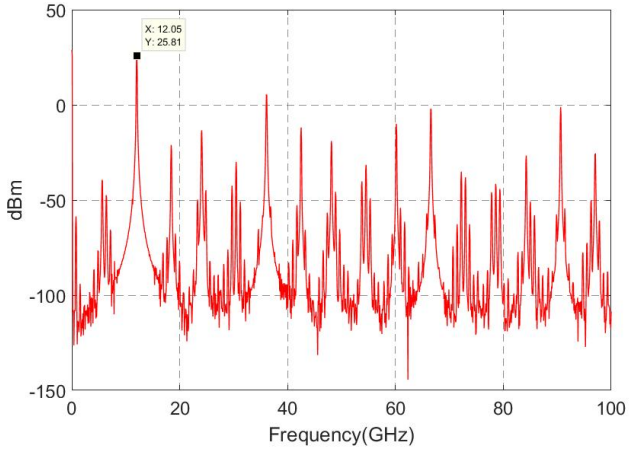


Figure 16: Typical power spectrum-wide band.

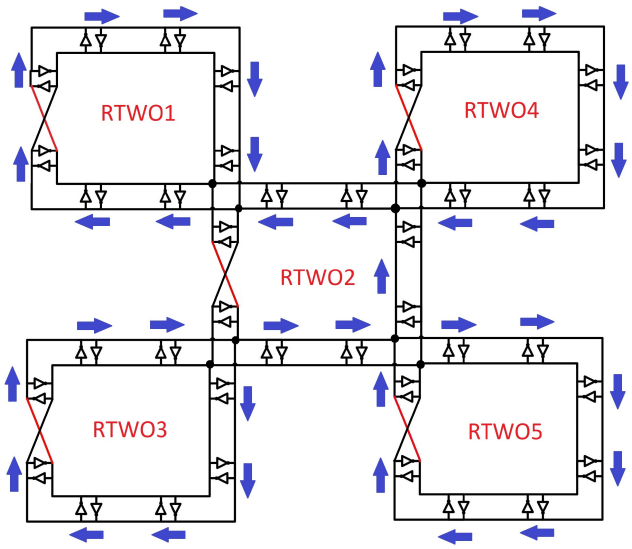


Figure 17: Diagram of the simulated ROA.

7 Design of an ROA

After designing a RTWO, an array of RTWO oscillators was made. In this case, a checkerboard configuration was used in which five RTWOs were connected (Figure 17). Each of the RTWOs is in phase with the others at the connection node. The advantage of a ROA is that multiphase signals can be obtained anywhere in the integrated circuit by reducing the skew and jitter of the signal.

The signal of each of the RTWOs of the ROA is shown in the Figure 18. In this case, only the signal of the external nodes of every RTWO is shown. It can be seen that the signals travel in clockwise direction, while in the middle RTWO (RTWO2) they travel in the opposite direction. Similarly, as the signal passes from one compensation stage to another, the signal is offset. For this case, the offset is given by Equation 5:

$$\Delta\Theta = \frac{180°}{\#stages} = 22.5° \tag{5}$$

Nevertheless, if small oscillation frequencies are desired, they must be made with maneuverable lengths so that L_t and C_{lt} increase their value and therefore the oscillation frequency decreases. This is a disadvantage in the design of RTWOs and ROAs because they will consume a larger area on the chip. Another way to decrease the frequency of oscillation is to be able to increase C_{comp}, which can be achieved by making larger transistors for the compensation stage. The disadvantage is that having very large transistors increases the power consumption, therefore there is a strong compromise between the area of the chip and the power consumption in low frequency RTWOs.

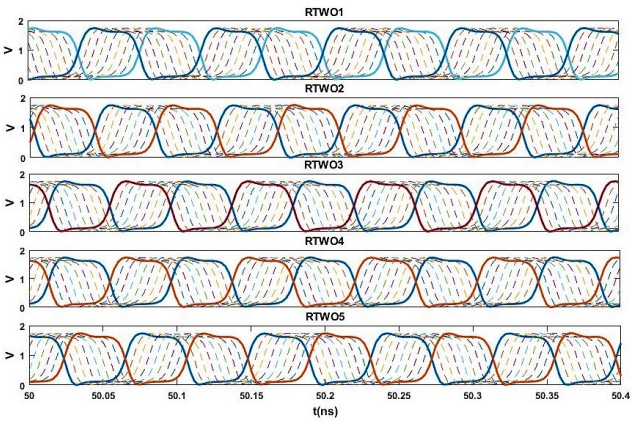


Figure 18: Simulated output waveforms.

8 RTWO performance comparison against existing works

In order to show the efficiency of the designed method applied at the present work, a comparison was made with two published methodologies: the methods of Chen et al. [11] and Aidoo et al. [15]. Chen et al. used analytical approximations in the determination

of the electrical parameters R, L and C of the RTWO, including different RTWOs at different frequencies simulated in HSPICE with parameter values of a technology 0.18μm SMIC and a 1.8 V power supply. In contrast, Aidoo et al. divided the RTWO structure into segments with equal dimensions while the parasite parameters were extracted from every segment using a 2D simulator. By this method of bonding segments, Aidoo et al. designed and manufactured different RTWOs using an IBM 0.18μm CMOS technology.

Using as reference the RTWO manufactured by Aidoo et al. at the present work (which oscillates up to 12.2 GHz), two RTWOs were designed with the same ring dimensions as well as of its compensators, but using 8 compensation stages for the design 1, and 28 for the design 2. Later, both the method used by Chen et al. and the one proposed in this work, were applied to extract the parasite parameters from the RTWO. The specifications of the different designs are shown in Table 4.

Since the comparison made in Table 4 considers RTWOs designed by parameters of different technological processes, this might not be a good comparison. In order to improve the comparison, two RTWOs were designed using the same technology (UMC 180 nm RFCMOS): one of them applying the method of Aiddo et al. and another one applying the method proposed in this work. Table 5 shows a comparison of the parasite parameters variations (R, L and C), the size of the transistors, the oscillation frequency and the amplitude of both designs. K_m represents the magnetic coupling coefficient of the ring differential line used to develop the electrical simulation in HSPICE. W_p and W_n are the dimensions of the transistors of the compensating circuit.

Table 4: RTWOs specifications using different manufacturing technologies (l= 880 μm, W =10 μm , S = 20 μm).

Parameter	Aidoo [15]	Chen [11]	Design 1	Design 2
$W_n(\mu m)$	48	52.1	52.1	15.3
$W_p(\mu m)$	115	140.6	140.6	41.4
Compensation stages number	8	8	8	28
f_{osc} simulated (GHz)	11.9	12.83	12.205	12.19
f_{osc} experimental (GHz)	12.2	—-	—-	—-
$f_{osc-sim}$- $f_{osc-exp}$ (%)	-2.45	5.16	0.04	-0.08

In Table 5, it is possible to observe that the oscillation frequency increases by 0.53% with respect to the one obtained by the method of Aidoo et al. This increase in the oscillation frequency agrees with the behavior between the simulated and the experimentally measured values of Aidoo et al. Moreover, the value of the resistance obtained from the simulation when applying the method of Aidoo et al. is lower than the extracted through the method developed at the present work. This is because the method of Aidoo et al. does not contemplate the proximity effect nor the skin effect on the return plane, and this variation in resistance directly affects the amplitude value. Therefore, the greatest number of electromagnetic effects must be considered to obtain more accurate results.

Table 5: Comparison between designed methods using the same technology.

Parameter	Aidoo [15]	This work	Variation (%)
$\Delta R(\Omega)$	0.231	0.286	23.78
$\Delta C(fF)$	15.753	14.720	-6.55
$\Delta L(pH)$	33.364	32.581	-2.34
$\Delta C_m(fF)$	1.548	2.252	45.45
K_m	0.072	0.052	-27.77
W_n (μm)	52.060	52.060	—-
W_p (μm)	140.580	140.580	—-
F_{osc} (GHz)	12.140	12.205	0.53
V_{out} (V)	1.770	1.685	-4.80

9 Conclusions

In this paper the impact of topology, dimensions and geometric discontinuities on power consumption and frequency of oscillation in an RTWO were investigated. To obtain RLC values with greater accuracy than other previous works [11, 19, 20, 17, 23] , the use of a 3D electromagnetic simulator was implemented. From the results presented, it is shown that the correct dimensioning of the resonator is vital in this type of oscillator. It was observed that at a frequency range of 11.4GHZ-11.5GHz it is possible to save up to 98% of the power consumption by making the transmission line with a W=5um. It is recommended to choose a W as small as possible where the resistive losses in the conductor allows the RTWO oscillate as it was designed.

Additionally, it is recommended to compensate the discontinuities using some mitered technique. According to the results obtained, it is possible to improve (increase) the frequency of oscillation up to 3.34% when changing the corners from 90° to corners at 45°. The simulations results show that the RTWO topology has a strong influence on the oscillation frequency and the power consumption, i.e. when the number of crossings in a lower metal increases, the oscillation frequency decreases and the power consumption of the resonator increases. Consequently, we conclude that it is better to elaborate the structure of the RTWO with only one crossing in a lower metal level.

References

[1] Samar Hussein, and Ahmed Eladawy, Design of a 85 GHz Rotary Traveling Wave Oscillator for Imaging Applications, International Journal on Recent and Innovation Trends in Computing and Communication,. vol. 4, no. 6, pp. 321 324, 2016.

[2] Paolo Maffezzoni, Bichoy Bahr, Zheng Zhang, and Luca Daniel, Oscillator Array Models for Associative Memory and Pattern Recognition, IEEE transactions on circuits and systemsregular papers, vol. 62, no. 6, pp. 1591-1598, June 2015.

[3] Omeed Momeni, and Ehsan Afshari, High Power Terahertz and Millimeter-Wave Oscillator Design: A Systematic Approach, IEEE Journal of solid-state circuits, vol. 46, no. 3, pp. 583- 597, March 2011

[4] A. J. Drake, K. J. Nowka, T. Y. Nguyen, J. L. Burns, and R. B. Brown, Resonant clocking using distributed parasitic capacitance, IEEE Journal of Solid-State Circuits, vol. 39, no. 9, pp. 1520 1528, Sept. 2004.

[5] J. Wood, T. Edwards, and S. Lipa, Rotary traveling-wave oscillator arrays: A new clock technology, IEEE J. Solid-State Circuits, vol. 36, no. 11, pp. 16541665, Nov. 2001.

[6] OMahony, F. ; Yue, C.P. ; Horowitz, M.A. ; Wong, S.S. A 10-GHz global clock distribution using coupled standing-wave oscillators, IEEE Journal of Solid-State Circuits, vol. 38, no. 11, pp. 1813- 1820, 2003.

[7] Zhengtao Yu, Xun Liu, Low-Power Rotary Clock Array Design, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, pp. 5 12, 2007.

[8] A. K. Rastogi, Munira Bano, Manisha Nigam, Microstrip Line Discontinuities Simulation at Microwave Frequencies, Advances in Physics Theories and Applications, vol.19, pp. 38-48, 2013.

[9] Sara Hosseini, Majid Shakeri, Sara Ebrahimian, Fatemeh Arastoonejad, A New Survey on Comparative Analysis of Discontinuities at Microstrip Lines Using Quasi-Static TLM SCN,. Journal of Basic and Applied Scientific Research, vol 3, no. 2, pp. 126-134, 2013.

[10] Koji Takinami, Rich Walsworth, Saleh Osman, and Steve Beccue , Phase-Noise Analysis in Rotary Traveling-Wave Oscillators Using Simple Physical Model, IEEE Transactions on microwave theory and techniques, vol. 58, no. 6, pp. 1465-1474, June 2010.

[11] L. Chen, H. Zhang, J. Zhou, and K. Chen, Efficient design of rotary traveling wave oscillator array via geometric programming, Journal of Zhejiang University-SCIENCE A, vol. 10, no. 12, pp. 18151823, 2009.

[12] C. Zhuo, H. Zhang, R. Samanta, J. Hu, and K. Chen, Modeling, optimization and control of rotary traveling-wave oscillator, in Proceedings of the 2007 IEEE/ACM international conference on Computer-aided design, pp. 476480, IEEE Press, 2007.

[13] Gonzalez, R. M. H., Aranda, M. L., Martínez, L. H., & Díaz, C. A. S. (2017, October), "Topology and discontinuities effect on CMOS rotary traveling wave oscillators". In Electrical Engineering, Computing Science and Automatic Control (CCE), 2017 14th International Conference on (pp. 1-5). IEEE.

[14] Y. Teng, J. Lu, and B. Taskin, Roa-brick topology for rotary resonant clocks, in Computer Design (ICCD), 2011 IEEE 29th International Conference on, pp. 273 278, IEEE, 2011.

[15] M. Aidoo, N. S. Dogan, Z. Xie, H. Savci, and P. Roblin, A power-aware rotary traveling wave oscillator (rtwo) design in 0.18 um cmos, in 2014 IEEE MTT-S International Microwave Symposium (IMS2014), pp. 14, IEEE, 2014.

[16] Chang, Richard Weng-Yew and See, Kye-Yak and Tan, Yang-Long, Impacts of bends and ground return vias on interconnects for high speed GHz designs, in Electromagnetic Compatibility and 19th International Zurich Symposium on Electromagnetic Compatibility, (APEMC2008), pp. 502-505, IEEE, 2008.

[17] L. Wu, R. Sebsebie, M. Kirkman-Bey, M. Aidoo, Z. Xie, and N. S. Dogan, Design of 110152 ghz rotary traveling wave oscillators in 65 nm cmos technology, in IEEE SOUTHEASTCON 2014, pp. 14, IEEE, 2014.

[18] Ma. Wei Fei, Hao Yu, and Junyan Ren, A 75.7 GHz to 102 GHz Rotary-traveling-wave VCO by tunable composite Right/left Hand Tline, IEEE Custom Integrated Circuits Conference, pp. , 2013.

[19] Armenta Márquez, Linares Aranda, and Hernández Martínez, Influence of discontinuities in Resonant Oscillators based on Transmission lines, 12th International Conference on Electrical Engineering, Computing Science and Automatic Control, pp. 593-598, 2015.

[20] M. Aidoo, CMOS Rotary TravelingWave Oscillators (RTWOs). PhD thesis, North Carolina AT State University, 2014.

[21] Raleigh Smith, Tuning of the Relative Phases of a Rotary Travelling Wave Oscillator, M. Sc. Thesis, 2014.

[22] Higuera Gonzalez, Caraterización de osciladores resonantes RTWO en tecnología CMOS, M. Sc. Thesis, 2017.

[23] Armenta Márquez, Oscilador Resonante de onda Rotatoria en Tecnologa CMOS, M. Sc. Thesis, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 265-280 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Parallelizing Combinatorial Optimization Heuristics with GPUs

Mohammad Harun Rashid*, Lixin Tao

Pace University, New York, USA

ARTICLE INFO

Article history:
Received: 12 August, 2018
Accepted: 09 November, 2018
Online: 18 November, 2018

Keywords:
GPU
Combinatorial
Optimization
Parallel
Heuristics

ABSTRACT

Combinatorial optimization problems are often NP-hard and too complex to be solved within a reasonable time frame by exact methods. Heuristic methods which do not offer a convergence guarantee could obtain some satisfactory resolution for combinatorial optimization problems. However, it is not only very time consuming for Central Processing Units (CPU) but also very difficult to obtain an optimized solution when solving large problem instances. So, parallelism can be a good technique for reducing the time complexity, as well as improving the solution quality. Nowadays Graphics Processing Units (GPUs) have evolved supporting general purpose computing. GPUs have become many core processors, multithreaded, highly parallel with high bandwidth memory and tremendous computational power due to the market demand for high definition and real time 3D graphics. Our proposed work aims to design an efficient GPU framework for parallelizing optimization heuristics by focusing on the followings: distribution of data processing efficiently between GPU and CPU, efficient memory management, efficient parallelism control. Our proposed GPU accelerated parallel models can be very efficient to parallelize heuristic methods for solving large scale combinatorial optimization problems. We have made a series of experiments with our proposed GPU framework to parallelize some heuristic methods such as simulated annealing, hill climbing, and genetic algorithm for solving combinatorial optimization problems like Graph Bisection problem, Travelling Salesman Problem (TSP). For performance evaluation, we've compared our experiment results with CPU based sequential solutions and all of our experimental evaluations show that parallelizing combinatorial optimization heuristics with our GPU framework provides with higher quality solutions within a reasonable time.

1. Introduction

Combinatorial optimization is a topic that consists of finding an optimal solution from a finite set of solutions. Combinatorial optimization problems are often NP hard. It is often time consuming and very complex for Central Processing Unit (CPU) to solve combinatorial optimization problems, especially when the problem is very large. Metaheuristic methods can help finding optimal solution within a reasonable time.

Nowadays efficient parallel metaheuristic methods have become an interesting and considerable topic. Local search metaheuristics (LSMs) are single solution-based approaches, as well as one of the most widely researched metaheuristics with various types such as simulated annealing (SA), hill climbing, iterated local search, tabu search, and genetic algorithm etc. A common feature that local search metaheuristics can share is, a neighborhood solution is selected iteratively as a candidate solution. We can use local search metaheuristics to solve combinatorial optimization problems such as graph bisection problem, Travelling Salesman Problem (TSP) etc. Many works have been done by using parallel computing technology to improve its performance from iteration level, algorithmic level and solution level. Therefore, parallelism is a good technique for improving the solution quality as well as reducing the time complexity.

Recently, Graphics Processing Units (GPUs) have been developed gradually to parallel/programmable processors from fixed function rendering devices. GPUs motivated by high definition 3D graphics from real time market demand have become many core processors, multithreaded, highly parallel with high bandwidth memory and tremendous computational power. So, more transistors are devoted to design GPU architecture in order to do more data processing than data caching/flow control. With the fast development of general purpose Graphics Processing Unit (GPGPU), some companies have promoted GPU programming

*Mohammad Harun Rashid, E-mail: harun7@yahoo.com

frameworks such as OpenCL (Open Computing Language), CUDA (Compute Unified Device Architecture), and direct Compute. GPU based metaheuristics have become much more computational efficient compared to CPU based metaheuristics. Furthermore, making local search algorithms optimized on GPU is an important problem for the maximum efficiency.

In the recent years, GPU computing has emerged as a very important challenge in the research areas for parallel computing. GPU computing is believed as an extremely useful technology for speeding up so many complex algorithms in order to improve solution quality. However, rethinking of existing parallel models as well as programming paradigms for allowing their deployment on GPU accelerators is one of the major challenges for metaheuristics. In fact the issue is, revisiting the parallel models as well as programming paradigms for efficiently considering the GPUs characteristics. However, some issues related to memory hierarchical management of GPU architecture need to be considered.

The contribution of this research is, to design a GPU framework which can efficiently deal with the following important challenges while parallelizing metaheuristics methods to solve large optimization problems:

1. Distribution of data processing between GPU and CPU with efficient CPU-GPU cooperation.
2. Thread synchronization and efficient parallelism control.
3. Data transfer optimization among various memories and memory capacity constraints with efficient memory management.

Our proposed parallel models on GPU architecture can be very useful and efficient in finding better optimized solution for large scale combinatorial optimization problems. We have made a series of experiments with our proposed GPU framework to parallelize some heuristic methods such as simulated annealing, hill climbing, genetic algorithm etc. for solving combinatorial optimization problems like Graph Bisection problem and Travelling Salesman Problem (TSP). All of our experimental evaluation shows that parallelizing heuristics methods with our GPU framework provides higher quality solutions in a reasonable computational time.

2. Background

As we contribute to the parallelization of heuristics methods for combinatorial optimization problems with GPU, below we discuss about combinatorial optimization problems, some optimization heuristics methods (hill climbing and simulated annealing), graph bisection problem/ travelling salesman problem as example optimization problems and GPU architecture/ computing.

2.1 Significance of Combinatorial Optimization

Combinatorial optimization can be defined as a mathematical discipline with the interplay between computer science and mathematics [1]. Very roughly, it deals with the problem of making optimal choices in huge discrete sets of alternatives. Combinatorial optimization is a way to search through a large number of possible solutions for finding the best solution from them. When the number of possible solutions is really too large and it is impractical to search through them, we can apply different techniques to narrow down the set and speed up the search.

Many combinatorial optimization problems are known as NP hard. That means, the time needed for solving a problem instance to optimality grows exponentially with the size of the problem in the worst case. Hence, these problems are easy to understand and describe, but very hard to solve. Even it is practically impossible to determine all possibilities for problems of moderate size in order to identify the optimum. Therefore, heuristic approaches are considered as the reasonable way to solve hard combinatorial optimization problems. Hence, the abilities of researchers to construct and parameterize heuristic algorithms strongly impact algorithmic performance in terms of computation times and solution quality.

2.2 Combinatorial Optimization Heuristics

To deal with combinatorial optimization problems, the goal is to finding such an optimal/optimized solution, which can minimize the given cost function. The cost of the algorithms can exponentially increase while the complexity of the search space is growing up, and this can make the search of a solution not feasible.

Finding a suboptimal solution within a reasonable time is another way to address these problems. In some cases, We might even find the optimal solution in some cases. These techniques can be divided into two main groups: heuristics and metaheuristics.

A heuristic is an algorithm that tries to find optimized solutions to complex combinatorial problems, but there is no guarantee for its success. Most of the heuristics are based on human perceptions, understanding the problem characteristics and experiments, but they are not based on fixed mathematical analysis. The heuristic value should be based on comparisons of performance among the competing heuristics. The most important metrics for performance are quality of a solution, as well as the running time. The implementation of heuristic algorithms is easy and they can find better solutions with relatively small computational effort. However, a heuristic algorithm can rarely find the best solution for large problems.

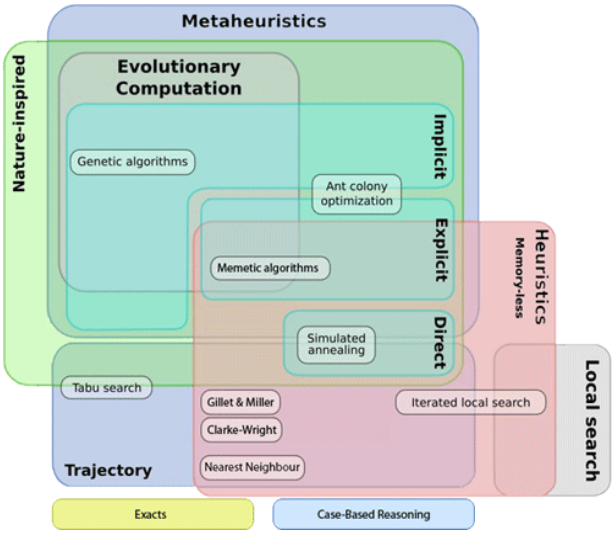


Figure 1: Different classifications of metaheuristics

Over the last couple of decades many researchers have been studying optimization heuristics for solving many real life NP hard problems, and some of the common problem solving techniques/methods underlying these heuristics came up as meta heuristics. A meta heuristic can be defined as a pattern or the major idea for a class of heuristics. Reusable knowledge in heuristic design can be represented by Meta-heuristics, which can provide us with important starting points in designing effective new heuristics for solving new NP hard problems.

Meta heuristics are not algorithms, nor based on theory. To effectively solve a meta heuristic based problem, we must have better understanding of the problem characteristics, and creatively designing as well as implementing the major meta heuristic components. So, it has become an action of research to use a meta heuristic for proposing an effective heuristic to solve an NP hard problem.

Another classification dimension is population based searches (P-Metaheuristics) vs single solution (S-Metaheuristics). P-Metaheuristics approaches maintain/ improve multiple candidate solutions, often using population characteristics to guide the search. P-Metaheuristics include genetic algorithms, evolutionary computation, and particle swarm optimization. S-Metaheuristics approaches focus on modifying/improving a single candidate solution. S-Metaheuristics include iterated local search, simulated annealing, hill climbing, variable neighborhood search etc. Below are some of the heuristics methods to solve combinatorial optimization problems:

2.2.1 Hill Climbing

Hill climbing is a mathematical optimization technique that belongs to the local search family and can be used for solving combinatorial optimization problems. The best use of this technique is in problems with "the property that the state description itself contains all the information needed for a solution". Hill climbing algorithm is memory efficient, since it does not maintain any search tree. This algorithm mainly looks into the present state and the immediate future states only. By using an evaluation function, it tries to improve the current state iteratively.

Hill climbing can encounter a problem called "local maxima". When the algorithm stops making progress towards an optimal solution, local maxima problem can occur because of the lack of immediate improvement in adjacent states. There are variety of methods to avoid Local maxima. Repeated explorations of the problem space could be one of the methods for solving this problem.

Hill Climbing Algorithm:

```
Get a random initial partition as the current partition.
While there is any improvement to the best cost seen so far
    Generate a random neighbor of the current partition.
    Evaluate the neighbor's cost.
    If the neighbor's cost improves the current cost
        Let the neighbor be the new current partition.
        If the neighbor's cost improves the best one seen so
far, record it.
    End If.
End While.
Return the best partition visited.
```

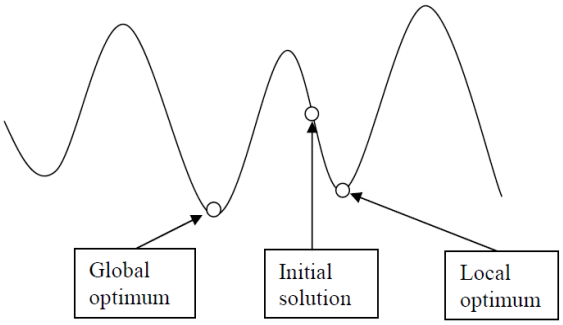


Figure 2: global solutions vs local solutions.

This algorithm starts from a random initial solution and then it keeps looking in the solution space in order to migrate to better neighbor solution. We might need to compare the current partition with all the neighbor's solutions before the algorithm is terminated. The algorithm terminates when all the neighbors' solutions are worse compared with the current partition. This technique can only find the local optimum solutions which are better solutions than all the neighbors, but the found solutions might not be global optimum solutions. The figure 2 shows difference between global and local solutions.

It is a time consuming process to maintain the visited neighbors of the current partition. That is why, it is necessary to parallelize the hill climbing algorithm for finding the best optimized solution.

2.2.2 Simulated Annealing

For graph bisection, simulated annealing heuristic starts with a high temperature t and a randomly selected initial partition as its current partition. After that, this heuristic starts the iterations with the same temperature and at each iteration, a neighbor partition is randomly generated. If the cost of the neighbor partition is better than the current cost, then the neighbor partition becomes the new current partition for the next iteration. If the neighbor partition does worsen the current cost, it can still be accepted with a probability as the new current partition. In case of high temperature, the probability is not sensitive to bad neighbor partition. But in case of low temperature, the probability for accepting a worsening neighbor will diminish with the extent of the worsening. The temperature is reduced by a very small amount after certain iterations are completed with the same temperature, and then, the iterations continue with the reduced temperature. The iteration process terminates once the termination criteria is met.

Many combinatorial optimization problems can be solved by applying simulated annealing heuristic. Unlike other meta heuristics, it has been mathematically proven that simulated annealing converges to the global optimum with sufficiently slow reduction of the temperature. As very few real world problems can afford such excessive execution time, this theoretical result does not interest much the practitioners. Below is the pseudocode for simulated annealing heuristic.

The simulated annealing and local optimization differ with the characteristics whether worsening neighbors will be accepted.

Simulated annealing heuristic starts with random walk in the solution space. If a random neighbor is better than the current solution, simulated annealing always accepts it. But, when the random neighbor is worse, the chance of accepting the worsen neighbor is slowly reduced. Simulated annealing can be reduced to local optimization with a very low temperature.

- Let $s = s_0$
- For $k = 0$ through k_{max} (exclusive):
  - $T \leftarrow$ temperature(k/k_{max})
  - Pick a random neighbour, $s_{new} \leftarrow$ neighbour(s)
  - If $P(E(s), E(s_{new}), T) \geq$ random$(0, 1)$:
    - $s \leftarrow s_{new}$
- Output: the final state s

Figure 3: Simulated Annealing Algorithm

2.3 Optimization Problems

2.3.1 Graph Bisection Problem

The graph bisection problem can be defined as a data representation of a graph G = (V,E) with a number of vertices=V and a number of edges=E, such that the graph G can be partitioned into smaller sections with some particular properties. For example, a k-way partition can divide the vertices into k smaller sections. When the number of edges between the separated components is relatively very small, it can be defined as a good partition. We can call a graph partitioning as uniform graph partition which divides the graph into smaller components in such a way that all the components are almost the same size as well as there are only small number of connections between the components. The important applications for graph partitioning include, but not limited to partitioning different stages for VLSI circuit design, scientific computing, clustering, task scheduling for multi-processing systems, and cliques detection in social networking etc.

Given a graph G = (V, E) where |V | is an even integer, find a partition of V into subsets L and R that minimizes the objective function

$$\text{cutSize}(L, R) = \sum_{(x,y)\in L\times R} adj(x, y)$$

under the constraint that |L| = |R|.

Here, adj() is the adjacency matrix and cutSize represents the cost for the bisection of the given graph G.

A partition satisfying the above conditions is called an optimal solution to the problem.

Example: The following graph has been partitioned into two equal-sized subsets by a dotted line, and the partition has a cut size of 2, which is the minimal cut size possible. This partition is therefore called optimal.

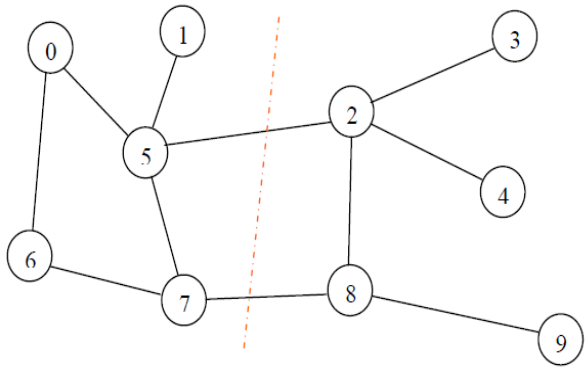


Figure 4: An optimal graph bisection

2.3.2 Travelling Salesman Problem

The Travelling Salesman Problem (TSP) is a combinatorial optimization problems which can be described easily, but it is very difficult to solve. A salesman starts with one city and will be visiting a number of cities with the condition that the salesman must visit each and every city only once and finally returns to first city. Selecting the sequence of the cities to be visited is the problem, because the salesman has to take the shortest path from a set of possible paths to minimize the path length. Exhaustive search can be used to find an optimal solution for a small instances (a few cities only) of TSP. But the problem is really critical for large number of cities, since with the increase in the number of cities, the number of possible paths increases exponentially. The number of possible paths for visiting n number of cities is the permutation of n which is n!. If the number of cities is increased by 1 only, the number of possible paths will become (n+1)!. Therefore, it will take too much time to compute the cost for all possible paths and find out the shortest path from them. TSP in known as a typical NP-hard problem.

TSP has a lot of applications in real world in different areas, like electronic maps, computer networking, Mailman's job, VLSI layout, traffic induction, electrical wiring, etc.

TSP Algorithms:

	TSP Greedy-Genetic Algorithm
Step 1:	Generate a random initial population of chromosomes (travel path).
Step 2:	Calculate fitness (path length) of every chromosome of initial population.
Step 3:	Repeat (iterations).
	a) Apply local search heuristic to select parents from initial population for cross over.
	b) Apply a greedy algorithm to generate greedy children from parents.
	c) Calculate fitness of all newly generated children and add newly generated children in initial population.
	d) Sort combined population initial population + new children by path length and select best chromosomes (travel path) for next population.
	e) Set initial population = next population.
	Until Terminating condition is satisfied (for given number of generations).

	Greedy-Children Algorithm
	input: an array `parentPath` containing n cities and an integer $k \leq n$:
Step 1:	Copy `parentPath`[1] into position 1 of a new array `childPath` of size n
Step 2:	Copy `parentPath`$[2, \ldots, k]$ into a new list `prefix`.
Step 3:	Copy `parentPath`$[k+1, \ldots, n]$ into `childPath`$[k+1, \ldots, n]$
Step 4:	Set `currentCity` = `parentPath`[1]
Step 5:	For $i = 2$ to k:
	a) Extract from `prefix` the city `nextCity` at minimum distance from `currentCity`.
	b) Store `nextCity` in `childPath`[i].
	c) Set `currentCity` = `nextCity`.
Step 6:	Return `childPath`.

2.4 GPU Architecture/Computing

Just a while ago, the conventional single core or multicore CPU processor was the only viable choice for parallel programming. Usually some of them were either loosely arranged as multicomputer in which the communication among them were done indirectly because of the isolated memory spaces, or tightly arranged as multiprocessors that shares a single memory space. CPU has a large cache as well as an important control unit, but it doesn't have many arithmetic logic units. CPU can manage different tasks in parallel which requires a lot of data, but data are stored in a cache for accelerating its accesses. Nowadays most of the personal computers have GPUs which offer a multithreaded, highly parallel and many core environments, and can potentially reduce the computational time. The performance of the modern GPU architecture is wonderful in regards to cost, power consumption, and occupied space.

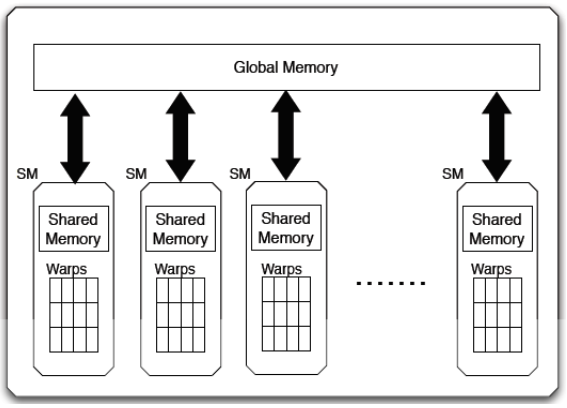


Figure 5: GPU architecture.

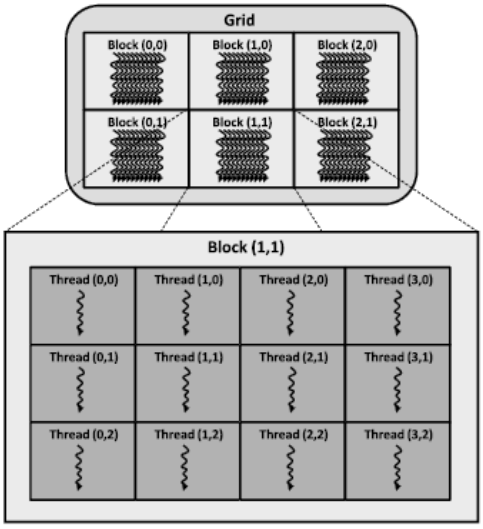


Figure 6: GPU thread blocks

A GPU includes a number of Streaming Multiprocessors (SMs). Each streaming multiprocessor contains a number of processing units which can execute thousands of operations concurrently. The warps inside a SM consist of a group of threads. A warp can execute 32 threads in a Single Instruction Multiple Data (SIMD) manner, which means all the threads in a warp can execute same operation on different data points. GPUs have at least two kinds of memory: global memory and shared memory. Global memory allows to store a large amount of data (such as 8GB), whereas shared memory can usually store only few Kilobytes per SM.

A GPU thread can be considered as a data element to be processed. GPU threads are very lightweight in comparison with CPU threads. So, it is not a costly operation when two threads change the context among each other. GPU threads are organized in blocks. Equally threaded multiple blocks execute a kernel. Each thread is assigned a unique id. The advantage for grouping of threaded blocks is that simultaneously processed blocks are linked closely to hardware resources. The threads within the same block are assigned to a single multiprocessor as a group. So, different multiprocessors are assigned to different threaded blocks. Therefore, controlling the threads parallelism can be a big issue for meeting memory constraints. As multiprocessors are mainly organized based on the Single Program Multiple Data (SPMD) model, the threads can access to different memory areas as well as can share the same code.

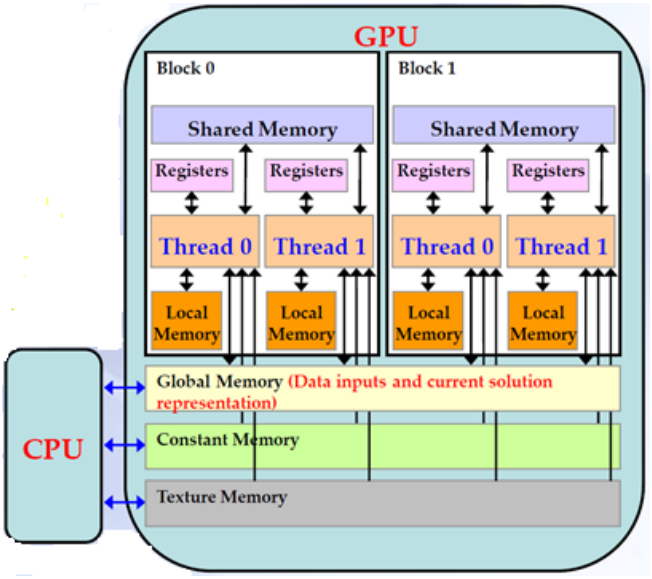


Figure 7: CPU - GPU communications.

GPU is used as a device coprocessor and CPU is used as a host. Each GPU has its own processing elements and memory which are separate from the host computer. Data is transferred between the host memory and the GPU memory during the execution of programs. Each device processor on GPU supports SPMD model, which means same program can simultaneously be executed on different data by multiple autonomous processors. To achieve this, we can define kernel concept. The kernel is basically a method or function which is executed by several processors simultaneously on the specified device in parallel and callable from the host as well. The Communications between CPU host and the device coprocessors are accomplished via the global memory.

3. Literature Review

In [2, 3, 4], the author proposed GPU based works with genetic algorithms. They proposed that the population evaluation as well as a specific mutation operator are to be performed in GPU. They implemented the selection and replacement operators in CPU. So, huge data transfers are performed between GPU and CPU. This kind of techniques can limit the performance of the solution.

In [5], the author proposed evolution strategy algorithm for solving continuous problems. According to his suggestions, multiple kernels can be designed for some of the evolutionary operators like selection, evaluation, crossover, mutation etc. and CPU can handle the rest of the search process. Later on, in [6], the author presented similar implementation with genetic algorithms. The additional contribution of their work was investigating the effect of problem size/ thread size /population size on GPU implementation comparing with sequential genetic algorithm.

In [7], the author proposed a memory management concept for an optimization problem. They implemented the concept for quadratic assignment problem where the global memory accesses were coalesced, the shared memory was used for storing as many individuals as possible, and the constant memory associated with matrices. For dealing with data transfers, their approach was a full parallelization based search process. In this regards, they divided the global genetic algorithm into multiple individual genetic algorithms, such that a thread block is represented by each sub population. Because of the poor management of data structures, the speed-ups obtained in their solution for combinatorial problems are not convincing.

In [8], the author proposed a framework of the automatic parallelization with GPU for the evaluation function. Only the evaluation function code need to be specified in their approach and the users don't need to know CUDA keywords. This approach allows evaluating the population on GPU in a transparent way. But, this strategy has some problems, such as it lacks flexibility because of transferring the data and nonoptimized memory accesses. In addition, the solution is limited to the problems where no data structure is required.

In [9], the author proposed an implementation of an evolutionary algorithm which is a GPU based full parallelization of the search process. Without any problem structures they implemented this approach to make an application for continuous and discrete problems. Later on, the authors also submitted an implementation of their algorithm with multi GPUs [10]. However, since there are some challenging issues of the context management such as global memories of two separate GPU, their implementation with multiple GPUs does not really provide with any significant performance advantages.

In [11], the author implemented a model for continuous optimization problems in which is very similar to the previous model. In this model, shared memory is used to store each sub population and organized based on ring topology. Although the speed-ups for the obtained solution are better compared with a sequential algorithm, the implementation of this model was dedicated to few continuous optimization problems only. As by considering the two previous models no general methods were outlined, in [12], the author made some investigation on the parallel island model on GPU. By involving different memory managements, they designed three parallelization strategies and were able to address some issues.

In [13], the author proposed a multi start tabu search algorithm and implemented to the TSP as well as the flow shop scheduling problem. The parallelization is performed on GPU by using shared libraries, and one tabu search associated with each thread process. However, this approach requires so many local search algorithms for covering the memory access latency and so, this type of parallelization is not much effective. In [14], the author proposed similar approach with CUDA. In this approach, the memory management for optimization structure is done in the global memory and they implemented this to the quadratic assignment problem. However, as one local search associated with each thread, the solution performance is limited to the size of instance.

For designing of multi-start algorithms, in [15], the author provided general methodology which are applicable to local search methods like simulated annealing, hill climbing, or tabu search. They also have contribution regarding the relationship between available memories and data mostly used for the algorithms. But, the application of the GPU accelerated multistart model is very limited, because it requires so many local search algorithms at run time to become effective. In [16], the author proposed a GPU based hybrid genetic algorithm. In their approach, they implemented an island model where a cellular genetic algorithm is represented by each population. Also, the mutation step in their hybrid genetic algorithm is followed by hill climbing algorithm. They performed their implementation for the maximum satisfiability problem.

According to the previous work, the hill climbing need to be integrated with the island model as per the investigation of the full parallelization. In this regard, in [17] the author proposed the redesign of GPU based hybrid evolutionary algorithms which performs a hybridization with a local search. Their focus was on different neighborhoods generation on GPU, correlating to each individual to be mutated in the evolutionary process. This kind of mechanism may guarantee more flexibility.

In [18], the author introduced a GPU accelerated multi objective evolutionary algorithm. In his approach, he implemented some of the multi objective algorithms on GPU, but not with the selection of non-dominated solutions. There are more works on P-metaheuristics for GPU parallelization are proposed. The parallelization strategies used for these implementations are similar to the prior techniques mentioned above. These works include particle swarm optimization [19, 20, 21], genetic

programming [22, 23, 24, 25] and other evolutionary computation techniques [26, 27, 28].

3.1 Research issues and contributions

Most of the approaches in the literature are mainly based on either iteration level or algorithmic level. In other words, the approaches are based on basically either the simultaneous execution of cooperative/independent algorithms, or GPU accelerated parallel evaluation of solutions. Regarding the cooperation between CPU and GPU, there are some implementations which also consider the GPU parallelization of other treatments such as selection/variation operators for evolutionary algorithms. We may argue on the validity of these choices, since an execution profiling may show that such treatments are negligible compared to the evaluation of solutions. As mentioned above, for reducing the data transfer between GPU and CPU, a full GPU parallelization of metaheuristics may also be performed. The original semantics of the metaheuristic are altered in this case to fit the GPU execution model.

Regarding the control of parallelism, a single thread with one solution are associated in most of the implementations. Besides, some of the cooperative algorithms associate one threads block with one sub population and may take advantage of the threads model. However, so far we've not found any work that has been investigated for managing parallelism of the threads efficiently to meet the memory constraints. The previous implementations may not be robust while dealing with large problem instances or a large set of solutions. In Chapter 4, we will show how an efficient control of thread may allow introducing fault tolerance mechanism for GPU applications.

For the memory management, so far there is no explicit efforts made in most of the implementations for memory access optimizations. For example, one of the most important elements for speeding up is memory coalescing, and additionally, could consider local memories for reducing non coalesced accesses. However, some of the authors proposed the simple way of using the shared memory to cache, but there is no performance improvement guarantee in those approaches. Some other authors also put some explicit efforts for handling optimization structures with the different memories, but still there is no general guideline/outline from those works. In fact, a lot of time, the associations of memory strictly depend on the target optimization problem such as small problem instances and/or no data inputs.

The contribution of this research is: to design a GPU framework with a set of efficient algorithms which can efficiently address the above mentioned challenges by parallelizing major metaheuristics for combinatorial optimization problems on the CPU-GPU architecture and also, to validate the solution quality with graph bisection problem as well as Travelling Salesman Problem (TSP).

4. Proposed Method and Contribution

Below are the research methodologies that we followed for proposing our methods and adding contributions to our GPU framework.

- Studying data processing distribution between CPU and GPU, and finding the challenges for efficient CPU-GPU cooperation.
- Studying thread synchronization, parallelism control on GPU threads, and finding the challenges for efficient parallelism control.
- Designing algorithms to optimizing data transfer between various memories and memory capacity constraints with efficient memory management.
- Developing parallel combinatorial optimization algorithms/frameworks for the CPU-GPU architecture to efficiently deal with the above mentioned GPU challenges.
- Validating the solution quality and efficiency of the proposed frameworks/algorithms relative to those of the best sequential meta-heuristics with extensive experimental design for graph bisection problem and TSP.

4.1 Difficulties in parallelizing optimization heuristics on GPU

Most of the time the performance of a parallel algorithm may depend on how well the communication structure of the target parallel system is matched with the communication structure of the algorithm. Nowadays one class of parallel processing systems consists of a number of processors, each with its own private memory. In addition, each of these processor memory pairs are connected to a small number of other pairs in a fixed topology. In these systems, if two processor-memory pairs must share data, a message is constructed and sent through the interconnection network. Such a message must be forwarded through one or more intermediate processors in the network. This forwarding introduces delay and hence reduces the amount of speedup achieved.

In science and industry, local search methods are heuristic methods for solving very large optimization problems. Even though these iterative methods can reduce the computational time significantly, the iterative process can still be costly when dealing with very large problem instances. Although local search algorithms can also reduce the computational complexity for the search process, still it is very time consuming for CPU in case of objective function calculations, especially when the search space size is too large. Therefore, instead of traditional CPUs the GPUs can be used to find efficient alternative solutions for calculations.

It is not straightforward to parallelize combinatorial optimization heuristics on GPU. It requires a lot of efforts at both design level and implementation level. We need to achieve few scientific challenges which are mostly related to the hierarchical memory management. The major challenges are: the CPU-GPU data processing with efficient distribution, synchronization of different threads, the data transfer optimization between different memories and their capacity constraints. Such challenges must be considered in redesigning of parallel metaheuristic models on GPU-CPU architectures for solving large optimization problems.

The following major challenges are identified for designing parallel combinatorial optimization algorithms efficiently on CPU-GPU architecture:

1. **CPU-GPU Cooperation:** It is important to optimize the data transfer between GPU and CPU to achieve the best performance. For efficient CPU-GPU cooperation, reparation of task must be defined in metaheuristics.
2. **Parallelism control on GPU threads:** In order to satisfy the memory constraints, it is important to apply the control of threads efficiently, since the order of the threads' execution is unknown for parallel multithreading in GPU computing. Also, it is important to define the mapping efficiently between each of the candidate solutions and a single GPU thread which is designated with a unique thread ID assigned at runtime.
3. **Management of different memories**: The performance optimization of GPU accelerated applications sometimes depend on data access optimization that includes the proper use of different GPU memory spaces. In this regard, it is important to consider the sizes and access latencies of different GPU memories for efficient placement of different optimization structures on different memories.

Below are the contributions of this research that address the challenges mentioned in section 4.1:

4.2 Efficient cooperation between GPU and CPU

An efficient GPU-CPU cooperation requires sharing the work as well as optimizing data transfer between two components.

4.2.1 Task repartition on GPU

The iteration level parallel model focuses on the parallelization of each iteration of metaheuristics. Indeed, the most time consuming task in a metaheuristic is the evaluation of the generated solutions. The concerns for the parallelization is the search techniques/mechanisms which are problem independent operations (For example, the evaluation of successive populations for P-metaheuristics and the generation/evaluation of the neighborhood for S-metaheuristics). As the iteration level model does not change heuristic's behavior, it can be defined as a low level Master Worker model. The following Figure 8 illustrates this Master Worker model. A set of solutions generated by the master at each iteration need to be evaluated. Each worker receives a partition of the solutions set from the master. The solutions are then evaluated by the worker and sent back to the master. In case of S-metaheuristics, the workers can generate the neighbors. Each worker receives the current solution from the master, generates neighbors for evaluation and then return this to master. This model is generic and reusable, since it is problem independent.

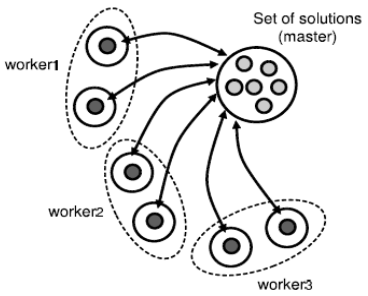


Figure 8: The parallel evaluation of iteration-level model.

As mentioned above, the most time consuming task of metaheuristics is often the evaluation of solution candidates. So, in regards with the iteration level parallel model the evaluation of solution candidates should be performed in parallel. According to the Master Worker model, the solutions can be evaluated in parallel with GPU. We can design the iteration level parallel model based on the data parallel SPMD (single program multiple data) model to achieve this. As showed in Figure 9, the main concept for GPU-CPU task partitioning is that CPU is responsible to host as well as execute the whole sequential part of the handled metaheuristic. On the other hand, the GPU is responsible for the solutions' evaluation at each iteration. The function code in this model called "kernel" to be executed on a number of GPU threads is sent to GPU. The number of threads per block determines the granularity of each partition.

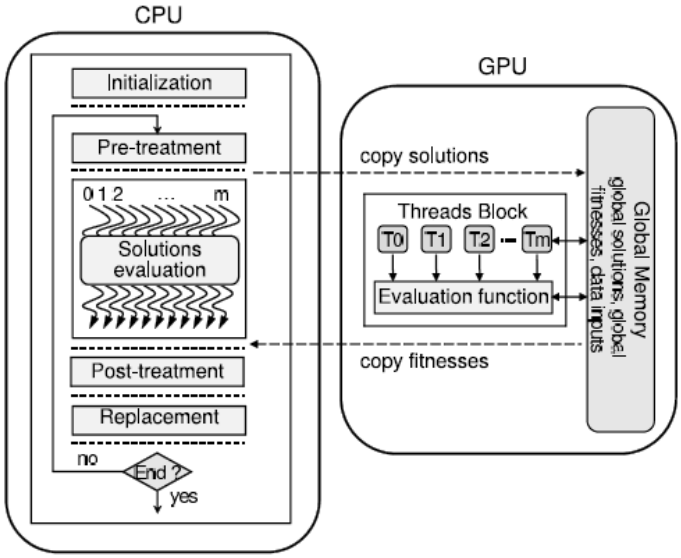


Figure 9: The parallel evaluation of solutions on GPU

4.2.2 Optimization of data transfer

Both GPU and the host computer have their own separate memories and processing elements. So, data transfer between GPU and CPU via PCI bus can be performance bottleneck for GPU applications. A higher volume of data to be copied while repeating the process thousands of times, definitely has a big impact on the execution time. For metaheuristics, the data to be copied are basically the solutions to be evaluated as well as their resulting fitnesses. For most of the P-metaheuristics, the solutions at hand are usually uncorrelated, but for S-metaheuristics each neighboring solution varies slightly compared to the initial candidate solution. So, for parallelization, the data transfers optimization is more prominent in case of S-metaheuristics.

In deterministic S-metaheuristics (such as hill climbing, variable neighborhood search, tabu search), the generation as well as evaluation of the neighborhood can be performed in parallel, which is indeed the most computation intensive. One challenge for data transfer optimization between GPU and CPU is to define where in S-metaheuristics the neighborhood should be generated. Below are two fundamental approaches for this challenge:

- **Neighborhood generation on CPU, but evaluation on GPU**: The neighborhood is generated on the CPU at each iteration of the search process, and the structure associated with this to store the solutions is then copied to GPU. It is pretty straightforward, as a neighbor representation of a thread is associated automatically with it. Usually this is something that can be done to parallelize P-metaheuristics with GPU. So, in this approach the data transfers are basically: 1) Copying neighbor solutions from CPU to GPU 2) Copying fitnesses structures from GPU to CPU.
- **Both neighborhood generation and evaluation on GPU**: The generation of neighborhood happens dynamically on GPU and thereby, there is no need to allocate any explicit structures. A little variation with the candidate solution that generates the neighborhood can be considered as a neighbor. So in this case, only the candidate solution is copied from CPU into GPU. The main advantage is that the data transfers is reduced drastically because only the resulting fitnesses structure need to be copied back from GPU to CPU, but the entire neighborhood does not need to be copied. However, this approach has a problem of determining the mapping between a thread and a neighbor which might be challenging in some cases.

Although the first approach is straightforward, implementing this in S-metaheuristics with GPU will require a large volume of data transfers in case of large neighborhood. This approach can be implemented in P-metaheuristics because the whole population is generally copied from the CPU to the GPU. The first approach might affect the performance because of the external bandwidth limitation. That is why, we consider the second approach i.e. both generation of neighborhood and evaluation on GPU.

The Proposed GPU accelerated Algorithm:

It is not a simple task to adapt traditional S-metaheuristics to GPU. We propose an algorithm 4.2.2 in a generic way to rethink S-metaheuristics on GPU. Memory allocations are made on GPU at the initial stage and also, data inputs as well as candidate solution are allocated initially.

Algorithm 4.2.2: GPU accelerated S-metaheuristic Template

```
1: Select an initial solution and also, evaluate this solution
2: Initialize specific variables if needed
3: Allocate problem data inputs, a solution, a neighborhood
fitness's structure and additional structures of solution on GPU
device memory,
4: Copy data inputs of the problem, a solution, and additional
structures of solution on GPU device memory
5: repeat
6:      for each neighbor on GPU in parallel
7:          Evaluate the candidate solution
8:          Add the resulting fitness into the
neighborhood fitness's structure
9:      end
10:     Copy neighborhood fitness's structure on CPU host
memory
11:     Specific strategy for solution selection on the
neighborhood fitness's structure
12:     Specific post-treatment
13:     Copy the chosen solution and additional structures of
solution on GPU device memory
14: until the stop criterion is satisfied
```

As previously said, heavy computations are required by GPUs with predictable accesses of memory. Hence, we need to allocate a structure to store the results for evaluating each neighborhood fitness's structure at different addresses. To facilitate the computation of neighbor evaluation, we can also allocate additional solution structures that are problem dependent. The data inputs of the problem, initial candidate solution and additional solution structures need to be copied onto GPU (line#4). The data inputs of the problem are read only structure which doesn't change at the time of all executions of the S-metaheuristic. So, during all the execution their associated memories are copied for only once. In the parallel iteration level, the neighboring solutions are evaluated and resulting fitnesses are then copied to the neighborhood fitnesses structure (line #6 to #9). The neighborhood fitnesses structure need to be copied into CPU host memory, since it is not defined in which order the candidate neighboring solutions are evaluated (line #10). Then, a particular strategy for the selection of the solution is implemented on the neighborhood fitness's structure (line #11): CPU explores the neighborhood fitnesses structure in a sequential way. Finally, the chosen solution and additional structures of solution are copied into GPU device memory (#13). This process repeats until some stop criteria is met.

4.2.3 Additional optimization of data transfer

In some S-metaheuristics, the selection criteria to find the best solution are based on maximal or minimal fitness. So, only one value (maximal fitness or minimal fitness) can merely be copied from GPU to CPU. However, it is not straightforward to find the appropriate maximal/minimal fitness's, as the read/write memory operations are performed asynchronously. The traditional parallel techniques that strongly suggests the global synchronization of hundreds of threads can decrease the performance drastically. Therefore, the techniques for the parallel reduction of each thread block should be adapted to address this issue. The following algorithm describes the techniques for the parallel reduction of each thread block.

Algorithm: parallel reduction techniques

Input Parameter: InputFitnesses on Global memory;

```
1: SharedMem[ThreadId] := InputFitnesses[id]
2: Synchronize locally

3: for n := NumOfThreadsPerBlock/2 ; n > 0; n := n / 2
4:          if ThreadId < n
5:                    SharedMem[ThreadId] :=
Compare(SharedMem[ThreadId], SharedMem[ThreadId + n])
6:                    Synchronize locally
7:          end if
8: end for
9: if ThreadId = 0
10:         OutputFitnesses[blockId]:= SharedMem[0]
11: end if
```

Output: OutputFitnesses

One element of input fitnesses from global memory is basically loaded into shared memory by each thread (line #1 and line #2). The array elements are compared by pairs at each iteration of the loop (line #3 to #7). As threads operate on different

memory addresses, the maximum/minimum of a given array can be found via the shared memory by applying local threads synchronizations in a given block. We can find the maximum or minimum fitness for all neighbors after a number of iterations are operated on GPU reduction kernel.

In some S-metaheuristics (such as simulated annealing), indeed the best neighbor is selected by the selection of maximal fitness or minimal fitness at each iteration. Therefore, the entire fitness's structure doesn't need to be transferred for these algorithms, and also, further optimizations might be possible.

4.3 Efficient control of parallelism

The efficient parallelism control on GPU for the iteration level is mainly focused here. For parallel multithreading in GPU computing, the order for the execution of threads is unknown, since it is indeed hyper threading based. First, it is important to apply the control of threads efficiently in order to satisfy the memory constraints. This allows to improve the overall performance by adding some robustness in the developed metaheuristics on GPU. Second, it is important for S-metaheuristics to define the mapping efficiently between GPU thread and each neighboring candidate solution. Therefore, at runtime each GPU thread is assigned with a unique thread ID for this purpose.

The key components for the parallelism control are the heuristic for controlling threads and the efficient mappings of the neighborhood structures. New S metaheuristics can be designed on GPU by considering these key components. The difficulty arises from the sequential characteristics of the metaheuristics that are first improvement based. In case of traditional parallel architectures, the neighborhood is generally divided into separate partitions with equal size. Then the generated partitions are evaluated and when an improved neighbor is found, the exploration stops. The whole neighborhood doesn't need to be explored, as the parallel model is asynchronous. When the computations become asynchronous, GPU computing is not efficient to execute such algorithms because GPUs' execution model is basically SIMD. Moreover, as the execution order of GPU threads is not defined, no such inherent mechanism exists for stopping the kernel in its execution.

We can deal with this type of asynchronous parallelization by transforming these algorithms into a data parallel regular application. So, we can consider the previous iteration-level parallelization scheme on GPU, which means that instead of applying to the entire neighborhood, we can apply to a sub set of solutions which need to be generated and evaluated on GPU. A specific post treatment on this partial set of solutions is performed on CPU after this parallel evaluation. This approach is considered as a parallel technic for simulating the first improvement based S-metaheuristic. Regarding implementation, this is similar to the Algorithm 4.2.2 that is proposed in the previous section. The sub set that has to be handled is the only difference concerns, in which the neighbors are randomly selected. The heuristic of thread control can adjust the remaining parameters automatically once the number of neighbors are set.

Although it may be normal to deal with such an asynchronous algorithm, but compared to an S-metaheuristic this approach may not be efficient, because a full neighborhood exploration is performed on GPU in case of S-metaheuristic. Indeed, memory accesses need to constitute an adjacent range of addresses to get coalesced in order to get a better global memory performance. But this cannot be achieved for exploring a partial neighborhood, because neighbors are chosen randomly.

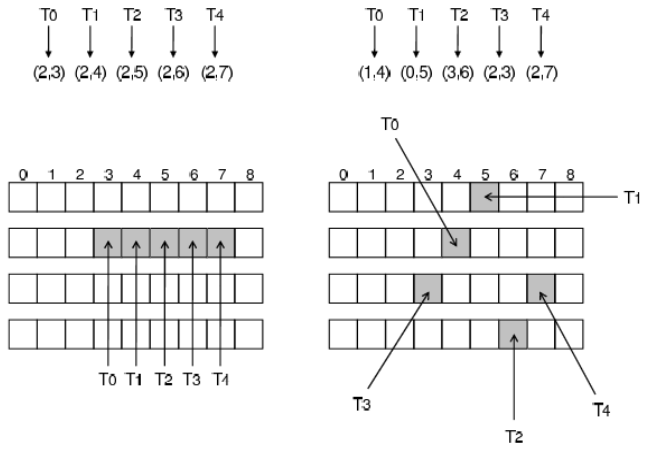


Figure 10: Illustration of a memory access pattern for both full exploration and partial exploration.

Figure 10 shows a memory access pattern for both full exploration and partial exploration of the neighborhood. In case of full exploration (left side of the Figure), all the neighbors are generated and many thread accesses get coalesced. In case of partial exploration of the neighborhood it does not happen (right side of the Figure), because no connection is available between the elements to get accessed.

4.4 Efficient memory management

For efficient implementation of parallel metaheuristics on GPU, it's important to understand the hierarchical organization of different memories. However, global memory coalescing can be done for some of the optimization structures which are specific to given GPU thread. This is usually the case for the large local structures that are used in P-metaheuristics for the evaluation function, or organization of data for a population.

Data accesses optimization that includes proper utilization of different memory spaces in GPU is important for optimizing the performance of GPU accelerated applications. The texture memory can certainly provide an amazing aggregation capabilities such as caching global memory. Each unit in texture memory gets some internal memory which buffers data from global memory. We can consider texture memory like a relaxed technique/mechanism of global memory access for the GPU threads, since coalescing is not required to accesses to this memory. For metaheuristics, the utilization of texture memory can be well adapted due to the following reasons:

- As no write operations are possible to perform on Texture memory, it is considered as a read only memory. This memory can be adapted in metaheuristics, because the inputs of the problem are read only values as well.
- In computation of evaluation methods, data accesses are frequent. The texture memory can make some high performance improvement by reducing the number of memory transactions
- In order to provide the best performance for 2D/1D access patterns, cached texture data is laid out. From a spatial locality perspective when the threads of a warp read locations are close together, then the best performance can be achieved. As the inputs of the optimization problems are generally 1D solution vectors or 2D matrices, we can bind the optimization structures to texture memory. Using of texture memories in place of global memory accesses is totally a mechanical transformation.

For parallelizing a metaheuristic, reducing the search time is one of the main goals and also, a fundamental aspect when there are some hard requirements on search time in some types of problems. In this regard, the parallel evaluation of solutions can be a concern for the iteration-level model. It can be considered as an acceleration model for the evaluation of independent as well as parallel computations. This is usually the case of S-metaheuristics that improve a single solution iteratively. There is no direct interaction between different neighborhood moves in these algorithms.

In case of P-metaheuristics, things are little different. During the search process, the solutions that represents a population can cooperate. For example, the solutions that compose the population are selected/reproduced by using variation operators in evolution based P-metaheuristics. A new solution can be constructed with different attributions of solutions which belong to the current population. Participating in constructing a common or shared structure (for example, ant colonies) is another example that concerns P-metaheuristics. The main input for generating the new population of solutions will be this shared structure, and the solutions that are generated previously participate in updating this type of common structure. Unlike S-metaheuristics, P-metaheuristics can provide additional cooperative aspects which is much more important while running multiple metaheuristics in parallel. The challenging issue here is the exploitation of these cooperative properties on GPU architectures.

To the best of our knowledge, these cooperative algorithms are never investigated much for CPU-GPU architecture. For P-metaheuristic, it is indeed the costliest operation to evaluate the fitness for each solution. Therefore, it is important to clearly define the task distribution in this scheme: for each cooperative algorithm the CPU is responsible for managing the whole sequential search process, whereas the GPU is responsible for evaluating the populations in parallel. The CPU sends a set of solutions through the global memory to be evaluated by GPU, and then, these solutions are processed on GPU. The same evaluation function kernel is executed on each GPU thread associated with one solution. Finally, the results of the evaluation function are sent to CPU through global memory.

Algorithm: GPU accelerated Cooperative algorithm for the parallel evaluation of populations

```
1: Select initial populations
2: Initialize specific variables if needed
3: Allocate problem data inputs, the different populations,
fitness's structures, additional structures of solution on GPU
4: Copy the problem data inputs to GPU
5: repeat
6:        for each P-metaheuristic
7:                particular pre-treatment
8:                Copy different populations as well as
additional structures of solution on GPU device memory

9:                for each solution on GPU in parallel
10:                        Evaluating Solution
11:                        Adding resulting fitnesses to
corresponding fitness's structure
12:                end for
13:                Copy fitness's structures on CPU (hosts
memory)
14:                particular post-treatment
15:                Population replacement
16:       end for
17:       Possible transfers between different P-metaheuristics
18: until some stop criteria is met
```

According to the above algorithm, memory allocations on GPU are made first i.e. problem data inputs, different populations and corresponding fitness's structures are allocated first (line #3). Additional structures of solution that are problem dependent can be allocated as well in order to make the computation of solution evaluation easier (line #3). Secondly, the data inputs of the problem need to be copied onto GPU device memory (line #4). The structure of these problem data inputs are read only, and also, for all the execution their associated memory need to be copied for one time only. Thirdly, the algorithm mainly describes that at each iteration different populations as well as the associated/ additional structures need to be copied (line #8). Then, the solutions are evaluated on GPU in parallel (lines #9 to #12). Fourthly, the structures of the finesses need to be copied into CPU (#13) and then a particular post treatment as well as population replacement are performed (line #14 and #15). Lastly, a possible migration can be performed on CPU at the end of each generation for information exchange between different P-metaheuristics (line #17). This process repeats until some stop criteria is met.

In this algorithm, GPU is utilized synchronously as a device coprocessor. However, as previously mentioned, copying operations (like population and finesses structures) from CPU to GPU can be a serious performance bottleneck. Accelerating the search process is the main goal of this scheme which does not alter the meaning of the algorithm. Hence, compared to the classic design on CPU, the policy of migration between the P-metaheuristics remains unchanged. This scheme is essentially

devoted to cooperative algorithms (synchronous), as the GPU is utilized as a device coprocessor for parallel evaluation of all individuals.

5. Experimental Validation

For our experimental validation, we have considered the following optimization problems to parallelize some heuristic methods with our proposed GPU framework in order to find higher quality solutions.

- Graph Bisection Problem
- Travelling Salesman Problem

5.1 Graph Bisection Problem

For Graph Bisection Problem, we have made experiments to parallelize hill climbing algorithm as well as simulated annealing algorithm with our GPU framework as follows:

5.1.1 Experimental Environment

OpenCL programming environment was setup on a NVIDIA CUDA GPU using C++ as follows:

- NVIDIA CUDA GPU (GeForce GTX 1050 Ti)
- OpenCL Driver for NVIDIA CUDA GPU
- VISUAL STUDIO 2017
- Windows 10 (64-bit Ultimate edition)

Created Visual Studio OpenCL projects (for both Hill Climbing and Simulated Annealing) using Visual C++.

5.1.2 Experimental Data

We've considered the following problem for the experiments:

Problem:

Prepare 200 GPU threads and run the heuristic method (Hill Climbing/Simulated Annealing) of graph bisection for the same problem instance to find the best solution/cost by considering the following factors:

- Each thread generates its initial random solution.
- Repeat 50 times to 100 times for each of the 200 threads to run heuristic method (Hill Climbing/Simulated Annealing) of graph bisection for the same problem instance.
- Keeps the running minimal cost and solution. They don't communicate.
- All the 200 threads pause. Find the smallest solution and its cost. Broadcast them to all the 200 threads as initial solution. Report the best solution and its cost.
- Many threads run in parallel. When the execution comes to a barrier method call, all threads suspend itself until all the threads have reached the barrier. Then they all do one operation (find minimum solution and broadcast to al threads), then resume their execution.
- Here, numbers 200, 50, and 100 are just random example numbers and we can change them based on the number of GPU threads that we can use.

Also for experimental validations, various problem instances (such as 20 vertices, 50 vertices, 100 vertices etc.) for graph bisections were used. An example of graph bisections problem for 20 vertices is given below:

0
1 1
0 0 1
0 1 0 0
0 0 0 1 0
0 1 0 1 0 0
1 0 0 0 0 1 0
0 1 0 1 0 0 0 1
1 0 0 0 0 0 1 1 0
0 1 0 0 0 0 1 0 1 0
0 0 0 0 0 0 0 1 0 0 0
0 0 0 0 0 1 1 1 0 0 0 1
0 0 0 0 1 0 1 1 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 1 0
1 1 0 0 1 1 0 0 0 1 0 1 1 1 0
1 0 1 0 0 1 0 1 0 1 0 0 0 1 0 1
0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1
0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0
0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 1 0 0 0

5.1.3 Experimental Data Presentation and Analysis

Below are the experimental presentation and analysis for both parallel Hill Climbing and simulated annealing algorithm with our GPU framework for Graph Bisection Problem. It is noted that we've used the following formula for calculating cut size as explained in section 2 (background).

$$\text{cutSize}(L, R) = \sum_{(x,y)\in L\times R} adj(x, y)$$

under the constraint that $|L| = |R|$.

Here, adj() is the adjacency matrix and cutSize represents the cost for the bisection of a given graph G = (V, E); where |V | is an even integer and we find the partition of V into subsets L and R that minimizes the objective function.

Data Presentation and Analysis for parallel Hill Climbing

As we ran both CPU based sequential hill climbing and our GPU based parallel hill climbing solution for multiple problem instances (For example: 20 vertices, 50 vertices, 100 vertices etc.) with the above mentioned scenarios and data, we've found the following results:

Table 1: Experiment Results for hill climbing algorithm on GPUs

Problem Instance	Best Cost for Sequential Solution	Best Cost for Parallel Solution	Improvement
Graph10.txt	4	3	25%
Graph15.txt	18	13	27.77%
Graph20.txt	25	18	28.00%
Graph30.txt	21	17	23.52%
Graph50.txt	18	15	16.66%
Graph100.txt	30	20	33.33%

We can see from the above results in Table 1 that for each problem instance the GPU based parallel solution got some good

improvement on cut size (cost) compared to CPU based sequential solution and thus, the average improvement on cut size (cost) for multiple problem instances is 25.71%. Therefore, much better optimized solution is found for Graph Bisection problem by parallelizing hill climbing algorithm on GPUs.

Data Presentation and Analysis for parallel Simulated Annealing

As we ran both CPU based sequential Simulated Annealing and GPU accelerated parallel Simulated Annealing solution for multiple problem instances ((For example: 30 vertices, 100 vertices etc.)) with the above mentioned scenarios and data, we've found the following results:

Table 2: Experiment Results for simulated annealing algorithm on GPUs.

Problem Instance	Best Cost for Sequential Solution	Best Cost for Parallel Solution	Improvement
Graph10.txt	5	5	0%
Graph15.txt	14	13	7.14%
Graph20.txt	19	15	21.05%
Graph30.txt	44	32	27.27%
Graph50.txt	133	60	54.88%
Graph100.txt	663	140	78.88%

We can see from the above results in Table 2 that for each problem instance the GPU based parallel solution got some good improvement on cut size (cost) compared to CPU based sequential solution. Thus, the average improvement for multiple problem instances on cut size (cost) is 31.53%. This improvement looks better when comparatively large problem instances are considered (such as: for 100X100 adjacent matrix, the improvement is 78.88%). Therefore, we can say that we've found a better optimal solution as we parallelize Simulated Annealing algorithm with GPUs.

5.2 Travelling Salesman Problem

5.2.1 Experiment Design

We built an experiment environment with the followings:

- CUDA programming model
- C++
- Visual Studio 2017
- NVIDIA GPU (GeForce GTX 1050 Ti)
- CPU (Core i7 9300 quad-core processor)
- Windows 10 (64-bit Ultimate edition)

For experimental validation of Travelling Salesman Problem (TSP), we've considered a CPU based sequential solution [29] that we previously proposed at an IEEE conference in 2017. For parallelization with our GPU framework, we've considered the following steps:

Step-1: Allocate 200 GPU threads.
Step-2: For each GPU threads.
a) Run our Main TSP Greedy-Genetic Algorithm in parallel to find the Path Length.

b) Send this thread obtained path length to CPU.

Step-3: CPU compares all the path lengths sent by all 200 GPU threads and determine the best path length.

It is noted that our Main TSP Greedy-Genetic algorithm that we previously implemented for a sequential solution [29] is illustrated in background section. We've parallelized the same proposed heuristics with our GPU framework, used the same data and compared the two results to confirm that we find the better optimized solution with GPU.

We developed our simulator by producing the inputs for Euclidean TSP and simplified the simulator with the following assumptions:

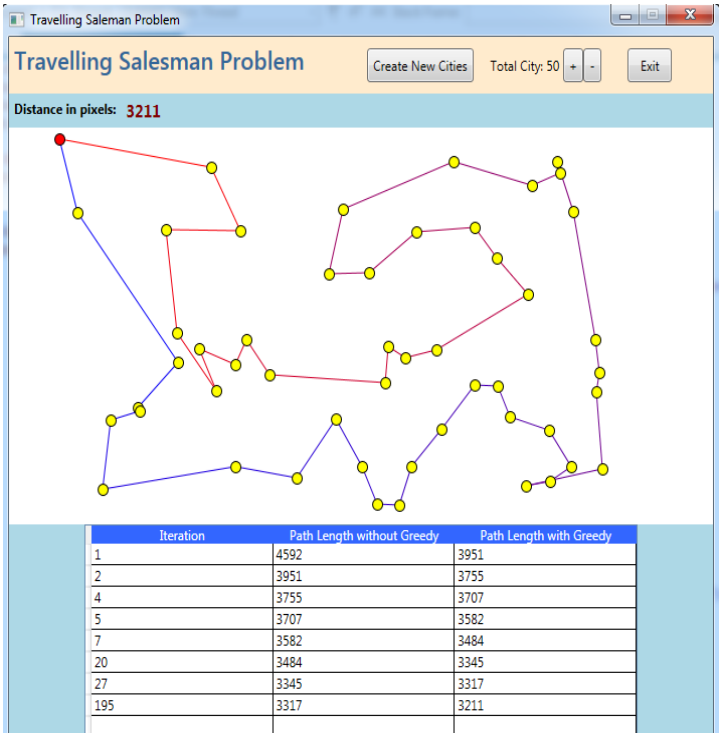
- The cities are located on the plane
- The distance between the cities is the Euclidean distance
- Each city is able to reach all other cities

We generated the inputs in such a way that the cities were uniformly placed on a grid at random with 600 columns and 350 rows. Then, by using the columns and rows as unit of Euclidean distance, the path length was obtained after so many numbers of iterations.

5.2.2 Experimental Data Tabulation and Visualization

We made 10+ repeated runs with the same instance in order to observe the behavior of our GPU accelerated parallel TSP solution. Table 3 shows the path lengths for n = 50 cities with 200 iterations. The input for each of this execution was generated randomly as explained above.

The following visual simulator developed for our previously proposed CPU based solution [29] shows the path length in different iteration levels which finally provide us with one minimum path length after completing all iterations.



We also randomly captured the following GPU thread results from our GPU based parallel solution for one single execution.

Table 3: Path lengths on different GPU threads for TSP

GPU Threads	Path length	Best Path length	Improvement
1	2086		
2	2019		
10	2162		
25	1895		
40	2362		
42	2108		
45	1965		
47	1888		
50	1919		
56	2169		
60	2009	**1776**	**12.68%**
65	2195		
68	2023		
75	2380		
85	1929		
88	1977		
100	1809		
115	1990		
160	1776		
200	2019		

After 10 times repeated run of our TSP heuristics on both GPU and CPU for the same number of cities (n=50), we obtained the following results:

Table 4: Comparison of path lengths between CPU and GPU for TSP

Run#	Path Length (CPU)	Path Length (GPU)	Improvement
1	2034	1776	12.68%
2	1965	1702	13.38%
3	2370	1904	19.66%
4	1881	1607	14.56%
5	2018	1780	11.79%
6	2243	1945	13.28%
7	1980	1756	11.31%
8	1777	1578	11.19%
9	2400	2019	15.87%
10	2018	1745	13.52%

5.2.3 Experiment Result Explanation

In our experimental validation, we involved 200 GPU threads to run the same TSP metaheuristics simultaneously by using our GPU framework. We captured different thread results which is illustrated in Table 4. We can see that the best path length calculated by thread 1 is 2086, thread 2 is 2019, ……..., thread 100 is 1809……, thread 200 is 2019 and so on. As all the thread results are sent to CPU for comparison, CPU finds the best path length (shortest) for TSP is 1776. The average path length obtained from different threads is 2034 and so, the improvement is 12.68% (approximately) because of the parallelization.

As we made 10 times repeated run of our TSP heuristics on both GPU and CPU for the same number of cities (n=50), both results are illustrated in Table 4. We can see that for run#3, CPU based best path length is 2370, whereas GPU best path length is 1904 which is much shorter that CPU based path length and thus the improvement on run#3 is 19.66%. Similarly, if we also observe the results for other repeated runs (#1, #2……...#10), we can easily notice that GPU based path length is definitely shorter than CPU based path length and thus there are some improvements in the solution quality for each execution on GPU.

The presented results show that the path lengths are shorter up to 19.66%, with an average of 13.72%. This improvement for finding the shortest path length in TSP is due to GPU parallelization with our framework.

6. Conclusion

In combinatorial optimization, parallel metaheuristic methods can be helpful to improve the effectiveness and robustness of a solution. But, their exploitation might make it possible to solve real world problems by only using important computational power. GPUs are based on high performance computing and it has been revealed that GPUs can provide such computational power. However, we have to consider that GPUs can have many issues related to memory hierarchical management, since parallel models' exploitation is not trivial. In this paper, a new guideline is established to design parallel meta heuristics and efficiently implement on GPU.

An efficient mapping of the GPU based iteration level parallel model is proposed. In the iteration level, CPU is used to manage the entire search process, whereas GPU is dedicated to work as a device coprocessor for intensive calculations. In our contributions, to achieve the best performance an efficient cooperation between CPU and GPU is very important because it minimizes the data transfer. Also, the goal for the parallelism control is, controlling the neighborhood generation to meet the memory constraints and also, finding the efficient mapping between the GPU threads and neighborhood solutions.

The redesigning of GPU based iteration level parallel model is suitable for most of the deterministic metaheuristics like Tabu search, Hill climbing, Simulated Annealing, or iterative local search. Moreover, we applied an efficient thread control to prove the robustness of our approach. This allows GPU accelerated metaheuristics preventing from crash when a large number of solutions are considered for evaluation. Also, this kind of thread control can provide some improvements with additional acceleration.

Redesigning of the algorithm for an efficient management of the memory on GPU is another contribution. Our contribution is basically the redesigning of GPU accelerated parallel metaheuristics. More specifically, we proposed multiple different general schemes to build efficient GPU based parallel metaheuristics as well as cooperative metaheuristics on GPU. In one scheme, the parallel evaluation of the population is combined with cooperative algorithms on GPU (iteration level). In regards to implementation, this approach is a very generic approach because we only considered the evaluation kernel. However, the performance is little limited in this approach because of data transfer between GPU and CPU. To address this issue, GPU based two other approaches operate on the complete distribution of search process, involving the appropriate use of local memories.

Applying such a strategy allows to extremely improve the performance. This approaches might experience some limitations because of the memory limitations with some of the problems which can be possibly more demanding with respect to resources. We have proved effectiveness of the proposed methods with a set of experiments in a general manner.

Furthermore, our experiments show that not only GPU computing exploits the parallelism to improve the solution quality, but also it can speed up the search process. In the future, we'll try to extend the framework with further features to be validated on a wider range of NP-hard problems in various fields like deep neural network, data science, artificial intelligence, computer vision, machine learning etc. including current industry challenges.

Conflict of Interest

We have no conflicts of interest to disclose.

References

[1] Dr. Lixin Tao, "Research Incubator: Combinatorial Optimization" Pace University, NY, February 2004

[2] Man Leung Wong, Tien-Tsin Wong, Ka-Ling Fok, "Parallel evolutionary algorithms on graphics processing unit" IEEE Congress on Evolutionary Computation, 2005. https://doi.org/10.1109/CEC.2005. 1554979

[3] Man-Leung Wong,Tien-Tsin Wong, "Parallel hybrid genetic algorithms on consumer-level graphics hardware" IEEE International Conference on Evolutionary Computation, 2006. https://doi.org/10.1109/CEC. 2006.1688683

[4] Ka-Ling Fok, Tien-Tsin Wong, Man-Leung Wong, "Evolutionary computing on consumer graphics hardware" IEEE Intelligent Systems, 22(2), 69–78, 2007. https://doi.org/10.1109/MIS.2007.28

[5] Weihang Zhu, "A study of parallel evolution strategy: pattern search on a gpu computing platform" GEC '09 Proceedings of the first ACM/SIGEVO Summit on Genetic and Evolutionary Computation, 765–772, 2009. https://doi.org/10.1145/1543834.1543939

[6] Ramnik Arora, Rupesh Tulshyan, Kalyanmoy Deb, "Parallelization of binary and real-coded genetic algorithms on gpu using cuda" IEEE Congress on Evolutionary Computation, 2010. https://doi.org/ 10.1109/CEC.2010.5586260

[7] Shigeyoshi Tsutsui, Noriyuki Fujimoto, "Solving quadratic assignment problems by genetic algorithms with gpu computation: a case study" GECCO '09 Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers, 2523–2530, 2009. https://doi.org/ 10.1145/1570256.1570355

[8] Ogier Maitre, Laurent A. Baumes, Nicolas Lachiche, Avelino Corma, Pierre Collet, "Coarse grain parallelization of evolutionary algorithms on gpgpu cards with easea" GECCO '09 Proceedings of the 11th An- nual conference on Genetic and evolutionary computation, 1403–1410, 2009. https://doi.org/10.1145/1569901.1570089

[9] Pablo Vidal, Enrique Alba, "Cellular genetic algorithm on graphic processing units" Springer Juan Gonz´alez, David Pelta, Carlos Cruz, Germ´an Terrazas, and Natalio Krasnogor, editors, Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), 2010. https://doi.org/10.1007/978-3-642-12538-6_19

[10] Pablo Vidal, Enrique Alba, "A multi-gpu implementation of a cellular genetic algorithm" IEEE Congress on Evolutionary Computation, 2010. https://doi.org/10.1109/CEC.2010.5586530

[11] Petr Pospichal, Jir´ı Jaros, Josef Schwarz. "Parallel genetic algorithm on the cuda architecture", In Cecilia Di Chio, Stefano Cagnoni, Carlos Cotta, Marc Ebner, Anik´o Ek´art, Anna Esparcia-Alc´azar, Chi Keong Goh, Juan J. Merelo Guerv´os, Ferrante Neri, Mike Preuss, Julian Togelius, and Georgios N. Yannakakis, editors, EvoApplications, Springer, 2010. https://doi.org/10.1007/978-3-642-12239-2_46

[12] Th´e Van Luong, Nouredine Melab, El-Ghazali Talbi. "GPU-based Island Model for Evolutionary Algorithms. Genetic and Evolutionary Computation Conference" GECCO '10 Proceedings of the 12th annual conference on Genetic and evolutionary computation, 1089–1096, ACM, 2010. https://doi.org/10.1145/1830483.1830685

[13] Adam Janiak, Wladyslaw A. Janiak, Maciej Lichtenstein. "Tabu search on gpu", J. UCS, 14(14):2416–2426, 2008.

[14] WZhu, J Curry, A Marquez. Simd, "tabu search with graphics hardware acceleration on the quadratic assignment problem" International Journal of Production Research, 2008.

[15] Th´e Van Luong, Nouredine Melab, El-Ghazali Talbi. "GPU-based Multi-start Local Search Algorithms" Coello C.A.C. (eds) Learning and Intelligent Optimization, Springer, 2011. https://doi.org/10.1007/978-3-642-25566-3_24

[16] Asim Munawar, Mohamed Wahib, Masaharu Munetomo, Kiyoshi Akama, "Hybrid of genetic algorithm and local search to solve maxsat problem using nvidia cuda framework." Genetic Programming and Evolvable Machines, 10(4), 391–415, 2009. https://doi.org/10.1007/ 978-3-642-25566-3_24

[17] Th´e Van Luong, Nouredine Melab, El-Ghazali Talbi, "Parallel Hybrid Evolutionary Algorithms on GPU" IEEE Congress on Evolutionary Computation, 2010. https://doi.org/ 10.1109/CEC.2010.5586403

[18] Man Leung Wong, "Parallel multi-objective evolutionary algorithms on graphics processing units" GECCO '09 Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers, 2515–2522, ACM, 2009. https://doi.org/10.1145/1570256.1570354

[19] Luca Mussi, Stefano Cagnoni, Fabio Daolio. "Gpu-based road sign detection using particle swarm optimization" IEEE Ninth International Conference on Intelligent Systems Design and Applications, 2009. https://doi.org/10.1109/ISDA.2009.88

[20] You Zhou, Ying Tan, "Gpu-based parallel particle swarm optimization" IEEE Congress on Evolutionary Computation, 2009. https://doi.org/ 10.1109/CEC.2009.4983119

[21] Boguslaw Rymut, Bogdan Kwolek, "Gpu-supported object tracking using adaptive appearance models and particle swarm optimization" In Leonard Bolc, Ryszard Tadeusiewicz, Leszek J. Chmielewski, and KonradW.Wojciechowski, editors, ICCVG, Springer, 2010. https://doi.org/10.1007/978-3-642-15907-7_28

[22] Simon Harding, Wolfgang Banzhaf, "Fast genetic programming on GPUs" In Proceedings of the 10th European Conference on Genetic Programming, 90–101. Springer, 2007.

[23] Darren M. Chitty, "A data parallel approach to genetic programming using programmable graphics hardware" GECCO '07 Proceedings of the 9th annual conference on Genetic and evolutionary computation, 1566–1573, ACM , 2007. https://doi.org/ 10.1145/1276958.1277274

[24] William B. Langdon, Wolfgang Banzhaf, "A SIMD interpreter for genetic programming on GPU graphics cards" EuroGP Proceedings of the 11th European Conference on Genetic Programming, 73–85, Springer,2008. https://doi.org/10.1007/978-3-540-78671-9_7

[25] William B. Langdon, "Graphics processing units and genetic programming: an overview. Soft Comput" Springer, 2011. https://doi.org/10.1007/s00500-011-0695-2

[26] Asim Munawar, Mohamed Wahib, Masaharu Munetomo, Kiyoshi Akama, "Theoretical and empirical analysis of a gpu based parallel bayesian optimization algorithm" IEEE International Conference on Parallel and Distributed Computing, Applications and Technologies, 2009. https://doi.org/10.1109/PDCAT.2009.32

[27] Lucas de P. Veronese, Renato "A. Krohling. Differential evolution algorithm on the gpu with c-cuda" IEEE Congress on Evolutionary Computation, 2010. https://doi.org/10.1109/CEC.2010.5586219

[28] Mar´ıa A. Franco, Natalio Krasnogor, Jaume Bacardit, "Speeding up the evaluation of evolutionary learning systems using gpgpus" GECCO '10 Proceedings of the 12th annual conference on Genetic and evolutionary computation,1039–1046, ACM, 2010. https://doi.org/10.1145/ 1830483.1830672

[29] Mohammad Rashid, Dr. Miguel A. Mosteiro, "A Greedy-Genetic Local-Search Heuristic for the Traveling Salesman Problem" IEEE ISPA/IUCC, 2017. https://doi.org/ 10.1109/ISPA/IUCC.2017.00132

Appendix:

Sample code for Kernel execution**:**

```
const char *kernelBytes = "\n" \
"__kernel void SimulatedAnnealing(
"    __global int* inBestPartition,
"    __global int* inBestCost,
"    __global int* outBestPartition,
"    __global int* outBestCost,
"    const int NumOfThreads)

"{
"    int gid = get_global_id(0);
"    int* FinalbestPartition;
"    int FinalBestCost;
"    barrier(CLK_GLOBAL_MEM_FENCE);
"    FinalbestPartition = inBestPartition[gid];
"    int newCost = inBestCost[gid];
"    if(newCost < FinalBestCost)
"    {
"        FinalBestCost = newCost;
"        FinalbestPartition = inBestPartition[gid];
"    }

"    if(gid < NumOfThreads)
"    {
"        outBestPartition[gid] = FinalbestPartition;
"        outBestCost[gid] = FinalBestCost;
"    }
"}
```

Sample code for parallel simulated annealing:

```
/*******************************************************
/*****                    Initialization of input data
/*******************************************************
const int vertexNumber = 100;
size_t sizeOfBuffers = vertexNumber*sizeof(int);
int* inputBestPartition = (int*)malloc(sizeOfBuffers);
int* inputBestCost = (int*)malloc(sizeOfBuffers);


//=======================SIMULATED ANNEALING==========
int bestPartition[vertexNumber];
randomPartition(bestPartition, vertexNumber);
int currentCost = cutSize(bestPartition, vertexNumber);
int bestCost = currentCost;                 // p[] is the

int neighbor[vertexNumber];   // Allocate space for a n
double t = 10.0;         // Initial temperature; parame
while (t > 0.01) {       // While not frozen; paramete
        copyArray(bestPartition, neighbor);
        randomSwap(neighbor, vertexNumber);         //
        int newCost = cutSize(neighbor, vertexNumber);
        int delta = newCost - currentCost;

        // Probability to accept a worsening neighbor
        double acceptProbability = exp(-delta / t);//
        // Otherwise take it with probability acceptPr

        if (delta <= 0) {
            // Accept the neighbor
            copyArray(neighbor, bestPartition);
            currentCost = newCost;

            // If the new solution is the best seen so
            if (currentCost < bestCost) {
                bestCost = currentCost;
            }
        }

    t = 0.95*t;   // Reduce temperature
}
```

Sample code for parallel hill climbing:

```
/*******************************************************
/*****                    Initialization of input data
/*******************************************************
const int vertexNumber = 20;
size_t sizeOfBuffers = vertexNumber*sizeof(int);
int* inputBestPartition = (int*)malloc(sizeOfBuffers);
int* inputBestCost = (int*)malloc(sizeOfBuffers);

//Initial hill climbing solution
int bestPartition[vertexNumber];
randomPartition(bestPartition,vertexNumber);
int bestCost =cutSize(bestPartition,vertexNumber);
//cout << "\nInitial Best Cost :" << bestCost << "\n" <<

// Loop terminates if we see 100 successive non-improving
int neighbor[vertexNumber];   // Allocate space for a nei
for (int i = 0; i < 100; i++) {
  copyArray(bestPartition, neighbor);
  randomSwap(neighbor,vertexNumber);      // neighbor[] :

  int newCost =cutSize(neighbor,vertexNumber);

  if (newCost < bestCost) {               // If new cost
    copyArray(neighbor, bestPartition);
    bestCost = newCost;

    i = 0;                                // retry 100 ti
  }
}

for(int i=0 ; i<vertexNumber ; i++)       //We put some r
{
    inputBestPartition[i] = bestPartition[i];
}

inputBestCost[0]=bestCost;
```

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 267-276 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Study on Improving Security and Efficiency using Dynamic Ownership Management in Client-Side Deduplication Environments

Won-Bin Kim, Im-Yeong Lee*

Department of Computer Science & Engineering, Soonchunhyang University, 31538, Republic of Korea

ARTICLE INFO

Article history:
Received: 21 August, 2018
Accepted: 24 September, 2018
Online: 12 October, 2018

Keywords:
Data Deduplication
Data Encryption
Cloud Storage

ABSTRACT

Data deduplication technology is used to improve the spatial efficiency of cloud storage. This technology is used for storing data on a cloud and omitting data uploading if the data are already present. However, various security threats may occur during the deduplication process. These security threats include poison attacks and user identity exposure through ownership. In addition, in an environment in which ownership changes in real time, there is a problem in renewing ownership information that has already been issued. Therefore, various studies have been conducted to solve these problems. In this study, a poison attack, real-time ownership management, and ownership anonymization are provided through MLE and dynamic ownership management.

1. Introduction

The rapid development of information communication and technology (ICT) in the recent times has led to many changes in data storage environment. According to a Dell EMC report in Korea, the amount of data produced in 2013 was 4.4 trillion GB, and this value is expected to increase to 44 trillion GB by 2020. As the amount of data produced increases, there is a demand for storage media with sufficient capacity to accommodate the data. Accordingly, in the future, the basic size unit of the storage medium will exceed terabytes and will be of the order of petabytes. However, portable hard disk drives (HDDs) and universal serial bus (USB) memories, which have been used in the past, have to be carried along always, and there exists the risk of losing them.

Cloud storage is a storage service that is available remotely over a network. It provides an environment wherein multiple users can access the storage simultaneously. Therefore, it is necessary to accommodate the data of a large number of users, and it involves maintenance and expansion costs, such as the costs for periodic storage space expansion, because it is necessary to ensure availability always. However, in general, much of the data stored by users are the same. Therefore, some of the storage space of the cloud storage is wasted in storing the same data repeatedly. To solve this problem, a data deduplication technique is proposed.

Data deduplication is a technique that reduces the amount of data stored in the data storage by preventing duplication of the stored data[1]. Because most of the data stored in the data storage are stored as the same data repeatedly, storage space is wasted. Data deduplication technology allows confirming that the data to be added are stored when the data are added to the storage. At this time, if the data to be added are already stored, the data are not stored, and the ownership of the data is given to the user. Therefore, the use of data deduplication technology can prevent repeated storage of the same data and improve storage space efficiency.

The cloud storage is a remote server. Therefore, data can always be leaked because of internal or external threats. The data stored in the storage must be encrypted so that the contents of the data can't be accessed by unauthorized users. However, common encryption techniques can't be applied simultaneously with deduplication because they don't know whether the two encrypted data originated from the same source. To solve this problem, a secure data deduplication technology using various technologies such as convergent encryption (CE) has been developed.

The initial technique for secure data deduplication was developed to enable simple deduplication of encrypted data. However, during the process of secure data deduplication, various security threats such as poison attack and ownership forgery attack occurred. In addition, a number of techniques were studied to solve such threats, but these techniques created additional problems such as the inconvenience of ownership management and excessive operation. In this paper, we propose an improved method of secure data deduplication. This paper is an extension of work originally presented in 2017 4th International Conference on Computer

*Corresponding Author. - Im-Yeong Lee, Email: imylee@sch.ac.kr

Applications and Information Processing Technology (CAIPT) [2].

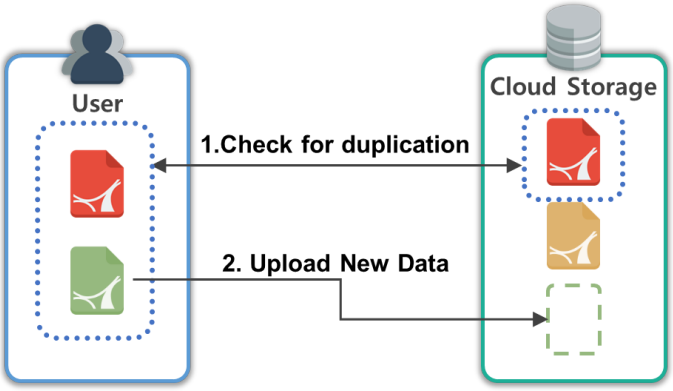


Figure 1. Data deduplication

2. Related Works

2.1. Data Deduplication

Data deduplication is a technique that prevents the same data from being repeatedly stored. To achieve this, when data are stored, it is necessary to check whether the same data are already present in the storage system or device as shown in Fig. 1. Therefore, with data deduplication a comparison of the data is conducted to check whether the same data are already stored. To do so, the data source is hashed and compared, and the data are stored according to the comparison results. Various methods can be applied during this process, and various types of systems can be designed according to the method used.

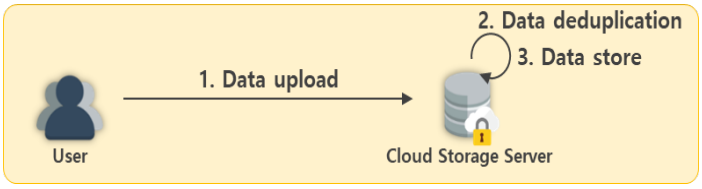


Figure 2. Server-side deduplication

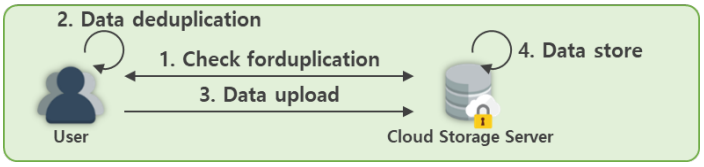


Figure 3. Client-side deduplication

2.2. Client-Side Deduplication (CSD)

Data deduplication uses a method for determining whether the data to be uploaded have already been stored. During this process, it is the responsibility of the storage server to determine whether the data already exist. However, the process of removing redundant data may take various forms. In the initial data deduplication method, all data to be uploaded are transmitted to the server, and the deduplication process is conducted in the server, as shown in Fig. 2. This is called server-side deduplication. However, because all data including redundant data are transmitted, a large amount of data transmission traffic occurs irrespective of the ratio of redundant data. In addition, bottlenecks may be incurred when data are uploaded from many different users

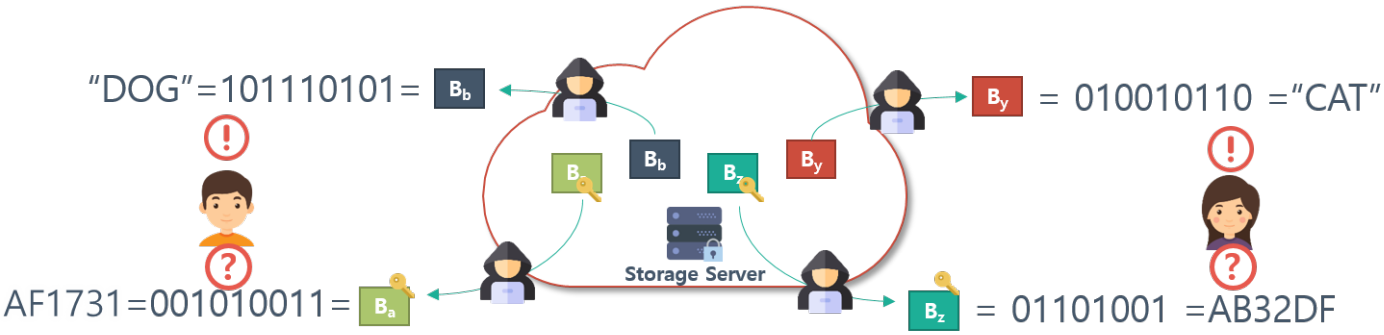


Figure 4. Necessity of data encryption

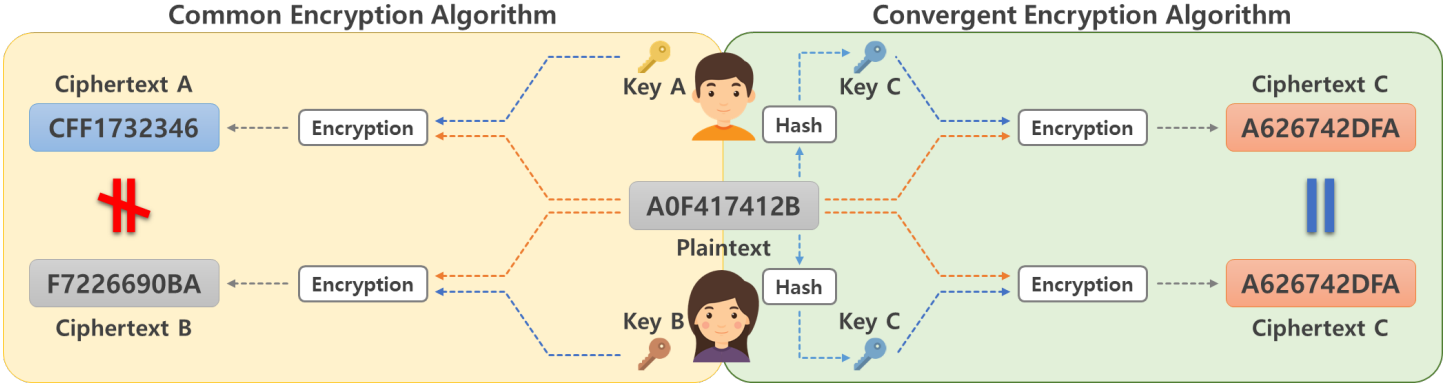


Figure 5. Comparison of common encryption and convergent encryption

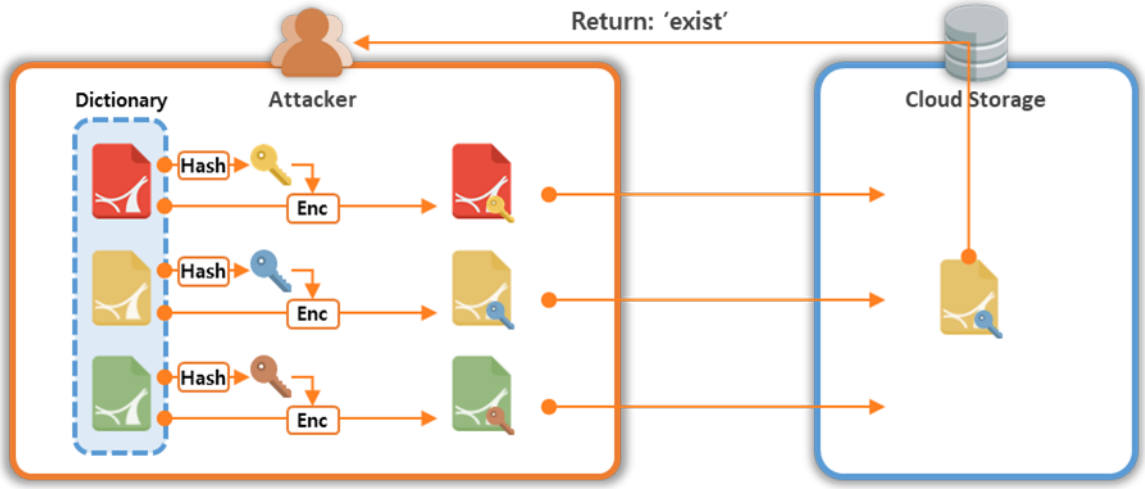


Figure 6. Dictionary attack scenario

concurrently. Therefore, client-side deduplication (CSD) technology has been proposed to solve this problem.

CSD is a method for determining redundancy by comparing a list of stored data with a list of identifiers from the data to be uploaded by the client to the server, as shown in Fig. 3. The server determines whether the data are redundant and transmits a list of non-duplicated data to the user. The user conducts deduplication using the duplicated data list received from the server, and transmits only the deduplicated data to the cloud storage server. CSD can reduce the amount of data transmission traffic and load on the server side. In this study, we used the CSD method [1-3].

2.3. Convergent Encryption (CE)

Data deduplication is conducted by comparing whether two data are the same. Therefore, we use hashed data for the data comparison. However, cloud storage is provided at a remote server, and thus has an honest-but-curious attribute. Therefore, threats continuously occur from insiders or from outside the system, and an encryption technique is thus required to prevent the data source from being known even when the data are leaked, as shown in Fig. 4. However, even if encrypted data are the same as the original data, the result depends on the encryption key used for encryption, and thus a data comparison based on the data deduplication technique cannot be conducted. To solve this problem, convergent encryption (CE), shown in Fig. 5, has been proposed [4].

CE is a technique for hashing a data source so as to generate hash data of a fixed size and using the source as an encryption key. Therefore, even if different users encrypt the data, the same encryption key and ciphertext are always generated. M. Storer proposed a technique for the deduplication of secure data using this approach [3]. First, key generation and encryption are conducted through the CE process. Thereafter, it is determined whether the data have been duplicated using the identifier of the encrypted data, and the data are additionally stored according to the result. Therefore, encryption and deduplication can be conducted concurrently. CE is used in most data deduplication environments, and various types of technologies have been proposed to utilize this technique. However, CE incurs a threat of poison attacks caused by different data and their identifier, as well as dictionary attacks that infer the data sources.

2.4. Dictionary Attack

A dictionary attack is an attack in which a ciphertext is decrypted, or related information is obtained, using pre-existing data. CE uses a method for hashing the data source to generate an encryption key, and encrypting it using the encryption key. Therefore, if an attacker knows the data source, the attacker can find the encrypted data and the encryption key, as shown in Fig. 6. Therefore, the attacker attempts an attack using a data list that is guessed as the data source. First, the attacker obtains the key by hashing the data that are guessed as the data source. When encryption is conducted using an encryption key and a data source, password data are generated. By comparing the generated secure data with a ciphertext maintained in storage, it is possible to confirm that the data generated by the attacker matches the original stored data. Such threats are quite serious in environments where data are predictable [5].

2.5. Poison Attack

A poison attack is a threat that occurs when the data stored do not match the actual data, as shown in Fig. 7 [6]. When a poison attack occurs in data M, two types of threats can be created:

- **Loss of data source:** If the user who acquired the ownership by performing a subsequent upload to the data where the poison attack occurred deletes the data source from the local storage, the original data can't be acquired again.
- **Damage due to malicious code or modified data:** Failure to detect tampering of data when downloading poison attack data may result in damage due to malicious code contained in the tampered data or property damage due to tampering with sensitive information.

To protect user's data from poison attacks, user need to be able to check whether stored data and metadata are generated from the same data. The RCE used in this study compares the tag with the data source and performs integrity verification after downloading

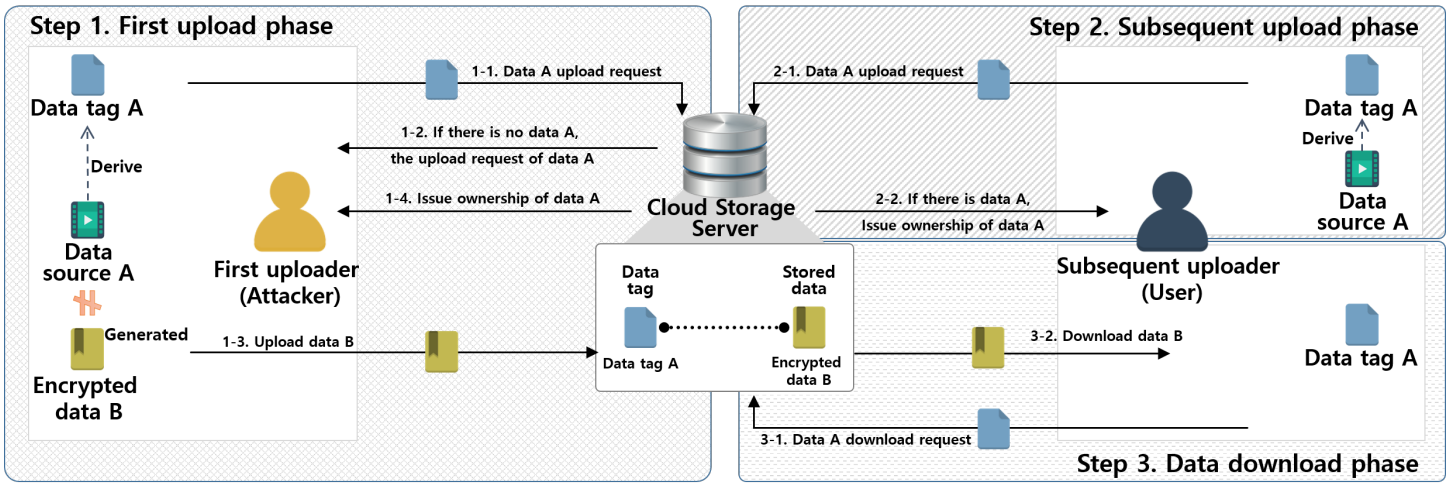


Figure 7. Poison attack scenario

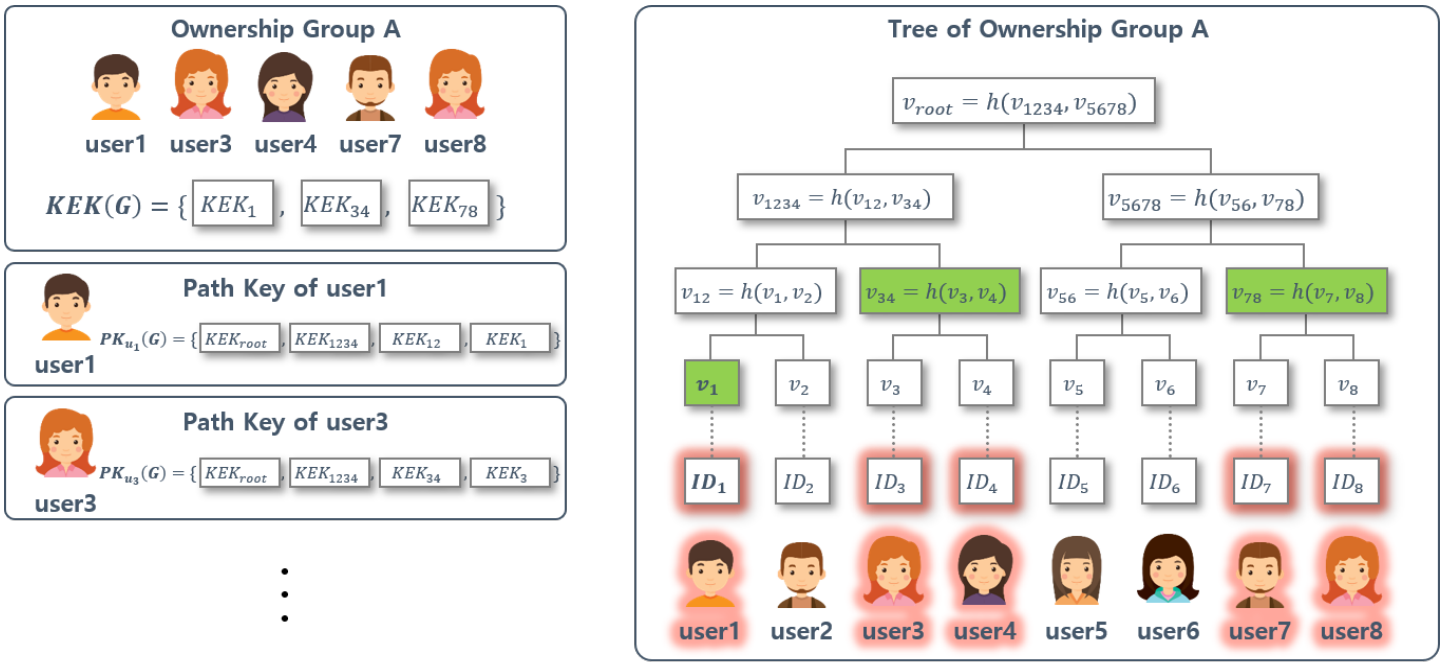


Figure. 8. Method of Ownership Management with Re-encryption

the data. However, since the direct association between the tag generated from M and the cipher text C can't be found in the data uploading stage, it is vulnerable to a poison attack.

2.6. Message-Locked Encryption

Message-locked encryption (MLE) offers four types of cryptographic techniques for the deduplication of encrypted data using technology proposed by Bellare et al. [7]. As a common MLE approach, data source M generates an encryption key K by computing K←H(M). Four types of MLE have been developed: CE, hash and CE without a tag check (HCE1), hash and CE with a tag check (HCE2), and randomized convergent encryption (RCE). The CE and HCE1 methods conduct only encryption and decryption. The HCE2 method conducts encryption and decryption along with tag integrity verification. The RCE method is based on HCE2, and the key generation entropy is improved through the generation of a randomized key in a one-time pad during the key generation process. In this study, a protocol design using RCE was applied.

2.7. Proof of Ownership (PoW)

In client-side deduplication technology, if the uploaded data are duplicated, ownership is issued without uploading the data. However, a problem arises if an attacker attempts to upload data by forging a data identifier. In this case, the attacker can take ownership of the data without the data source. To solve this problem, a technique is proposed to verify data ownership without transferring the data source. This technique is referred to as proof of ownership (PoW) [8].

2.8. Dynamic Ownership Management

A variety of data are stored on a cloud, and each data group has its own ownership group. However, if a change occurs, such as the issuing and discarding of ownership, a problem arises in that all

ownership of the ownership group must be renewed. In general, it is relatively easy to issue ownership to a new user. However, if ownership information for a particular portion of data changes, it is difficult to change the ownership previously issued to the user. Therefore, dynamic ownership management technology is necessary to manage changes in ownership information in real time. In this study, we use proxy re-encryption for dynamic ownership management.

2.9. Hur et al.'s scheme

Hur et al.'s scheme was proposed in 2016 [9]. The authors proposed a method for achieving dynamic ownership management in a secure data deduplication environment. This scheme addresses the problem of identifying the data owner using the ownership information of the user by providing anonymity of ownership. As a way to provide anonymity of ownership, a general method for confirming ownership through the ownership group was developed. However, this approach complicates the ownership management because the ownership of the entire group must be changed at the time of ownership issue and renewal. re-encryption was proposed to solve this problem, along with a dynamic ownership management technology providing better efficiency. However, this scheme is based on a server-side deduplication environment, is continuously affected by the redundancy rate of the data, and always involves the same number of computations. Therefore, in the present study, we researched an improved technology by applying the idea of Hur et al.'s scheme.

In Her et al.'s scheme, ownership management is performed using the Merkle Hash Tree (MHT) for dynamic management of data ownership. In the leaf node of the MHT, the identification information of the users is located, and the MHT is configured as shown in Fig. 8 using this identification information. Ownership group G_i contains a list of users having ownership of data M_i. Also, $KEK(G_i)$ is configured as shown in the left side of Fig. 8 so that users included in G_i can be included in the minimum number of nodes. Therefore, user1, user3, user4, user7, user8 belong to group G and $KEK(G)$ consists of KEK_1, KEK_{34} and KEK_{78} in Fig. 8. Also, for each user, a node included in the user's own identifier path from the root node is provided as a path key(PK). In Fig. 8, PK_{u_1} (G) is PK of user1. And it includes KEK_{root}, KEK_{1234}, KEK_{12}, and KEK_1.

The ownership manager generates $KEK(G_i)$, which is a list of KEK of the ownership group G_i, and provides $PK_{u_1}(G_i)$ to the user. Then, after generating the group key GK_i of the group G_i, the data M_i is encrypted to generate the cipher text $C_i = E_{GK_i}(M_i)$. Finally, the group key GK_i is encrypted with $KEK(G_i)$ to complete the update of the ownership group for the data M_i.

Through the above process, the user belonging to the ownership group can obtain the data $M_i = D_{PK_{u_1}(G_i) \cap KEK(G_i)}(C_i)$ by using the $PK_{u_1}(G_i)$ held by the ownership group and the $KEK(G_i)$ provided by the ownership manager.

3. System Requirements

3.1. System configuration

The system structure required in this study is shown in Fig. 9. Users access the cloud storage, which includes storage and metadata servers. Therefore, the user stores the metadata server, data tag, and verification data, and the encrypted data are stored in the storage server.

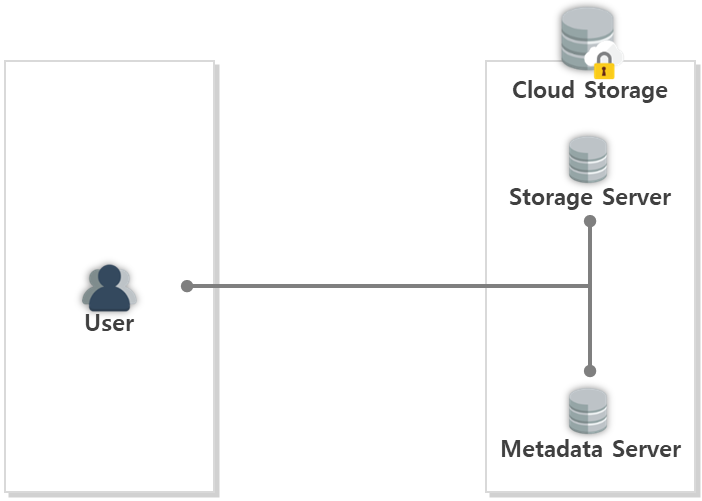


Figure 9. System configuration

3.2. Security requirements

The six security requirements are poison attack resistance, dictionary attack resistance, confidentiality, integrity, anonymity, and efficiency.

- **Confidentiality:** Data stored in cloud storage must be encrypted and archived in case of data leakage. Encryption should also be considered to enable data deduplication.
- **Integrity:** Data stored in cloud storage should be preserved without modification. If the data stored in the cloud storage are modified, the user should be able to detect the deformation when downloading the data.
- **Anonymity :** Ownership information stored in the server can only be used to verify the user's ownership, and information that can identify the user's identity should not be included.
- **Efficiency:** In the case of data deduplication, a large number of operations and data transmissions can occur. In this case, the computation and data traffic efficiency must be provided during the process because the benefits of the data deduplication are offset.
- **Resistance to poison attack:** A tag is derived from the data source and must be able to verify the integrity of the original data. This prevents poison attacks that may occur owing to the different tag sources and data stored in the server.
- **Resistance to dictionary attack:** An attacker should not be able to attack a dictionary. Therefore, an attacker should not be able to derive data or data encryption keys through the CE using an analogy with the data source.

4. Proposed Scheme

The proposed technique is based on Hur et al.'s concept of dynamic ownership management. In the proposed method, the data upload request phase is initially applied, and the first and subsequent data upload phases are conducted depending on whether the data are duplicated. If a poison attack is detected in the

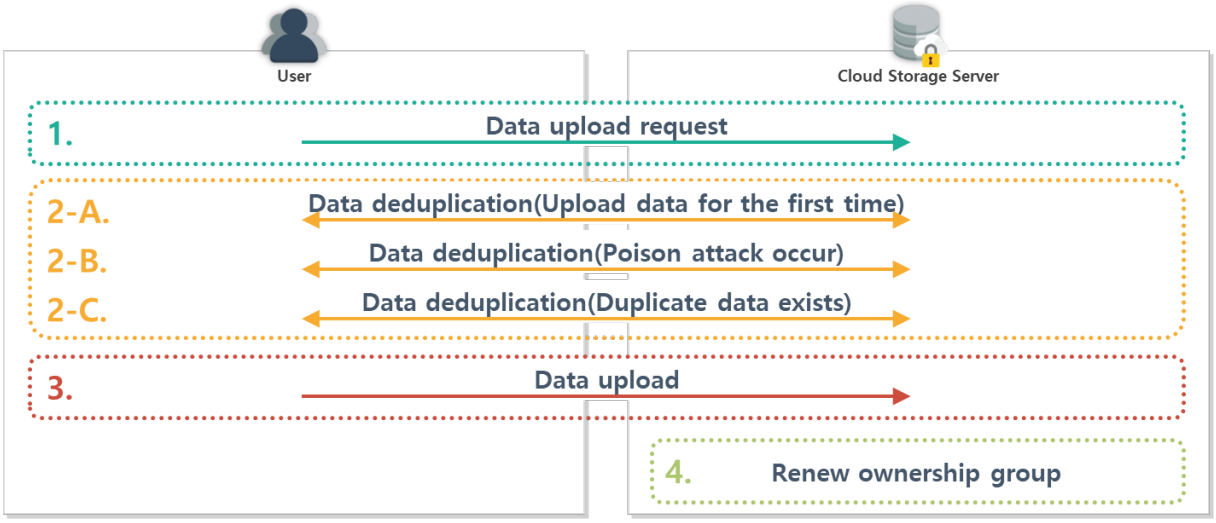


Figure 10. Overview of uploading data in the proposed scheme

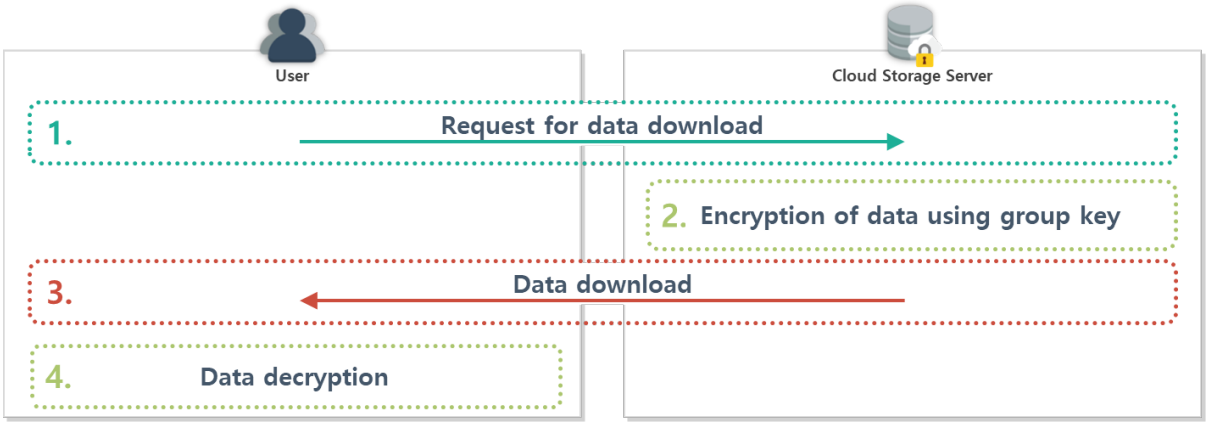


Figure 11. Overview of downloading data in the proposed scheme

$$K_i \leftarrow H(M_i) \tag{1}$$

4.1. Data Upload Request Phase

This step is a common step in the data upload process. At this stage, the user uses the identifier of the data to be uploaded, and requests the server to confirm whether the corresponding data exists, as shown in Fig. 12.

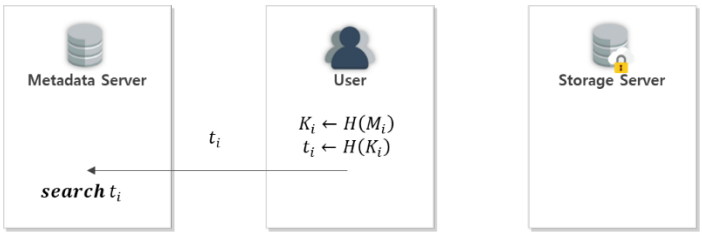


Figure 12. Data upload request phase

The user converts data M_i to be uploaded into a hash algorithm for generating K_i and t_i. The generated t_i is then transmitted to the metadata server to request confirmation of storage.

4.2. Data Upload Phase (First Upload)

This step is conducted when the corresponding data do not exist as a result of the data upload request phase, as shown in Fig. 13.

The metadata server determines that there are no data that a user has requested to upload and transmits the result to the user.

A user who receives a response from the metadata server generates L_i, C_i^1, C_i^2, and C_i^3.

$$L_i \xleftarrow{\$} \{0,1\}^{\lambda(K)} \tag{3}$$

$$C_i^1 \leftarrow E_{L_i}(M_i) \tag{4}$$

$$C_i^2 \leftarrow L_i \oplus K_i \tag{5}$$

$$C_i^3 \leftarrow H(M_i \parallel L_i) \tag{6}$$

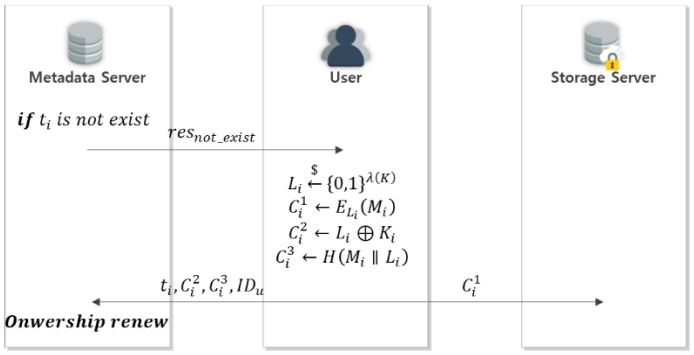


Figure 13. Data Upload Phase (First Upload)

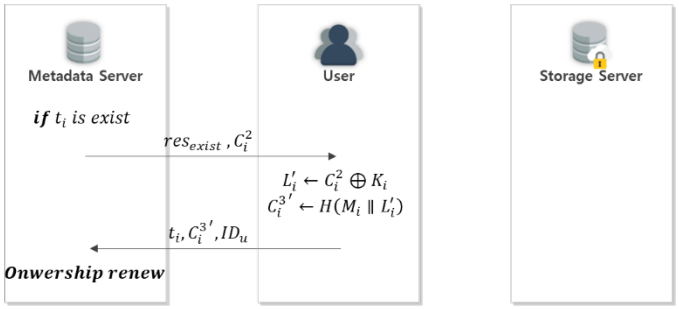


Figure 14. Data Upload Phase (Subsequent Upload)

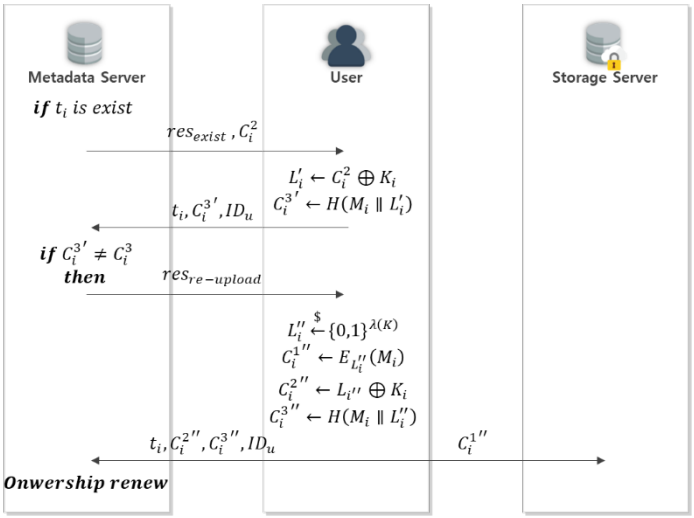


Figure 15. Data Upload Phase (when a Poison attack occur)

The user obtains L_i by calculating data C_i^2 obtained from the metadata server. Then, $C_i^{3'}$ is generated using the obtained L_i.

$$L_i' \leftarrow C_i^2 \oplus K_i \tag{7}$$

$$C_i^{3'} \leftarrow H(M_i \| L_i') \tag{8}$$

The user sends the created t_i, C_i^3 and ID_u to the metadata server. The ownership group renewal is then conducted.

4.4. Data Upload Phase (when a Poison attack occur)

This step is conducted when it is determined that a poison attack has occurred on the data uploaded during the data upload phase (subsequent upload), as shown in Fig. 15.

The metadata server determines that there are data that the user has requested to upload, and transmits the result and C_i^2 to the user.

The user obtains L_i by calculating data C_i^2 obtained from the metadata server, as shown in (7). Then, $C_i^{3'}$ is generated using the obtained L_i, as shown in (8).

The user sends the created t_i, C_i^3, and ID_u to the metadata server.

The metadata server determines that a poison attack has occurred when the user-uploaded $C_i^{3'}$ is different from C_i^3 stored in the metadata server.

The metadata server then requests the user to re-upload the data. Ownership group renewal is then conducted. A user who receives a response from the metadata server generates L_i'', $C_i^{1''}$, $C_i^{2''}$, and $C_i^{3''}$.

$$L_i'' \xleftarrow{\$} \{0,1\}^{\lambda(K)} \tag{9}$$

$$C_i^{1''} \leftarrow E_{L_i''}(M_i) \tag{10}$$

$$C_i^{2''} \leftarrow L_i'' \oplus K_i \tag{11}$$

$$C_i^{3''} \leftarrow H(M_i \| L_i'') \tag{12}$$

The user sends the generated t_i, $C_i^{2''}$, $C_i^{3''}$ and ID_u to the metadata server and transmits $C_i^{1''}$ to the storage server. Then, the ownership group renewal is conducted.

4.5. Ownership Group Renewal Phase

This step is executed after the data upload phase ends. Include the user who has been issued ownership in the ownership group. In this process, the metadata server encrypts and archives the GK_i using the added KEK. Fig. 16 shows the formation and management of the user's tree and ownership group.

The owner group G_i of data M_i has a key $KEK(G_i)$ for encrypting the group key GK_i. The $KEK(G_i)$ term is a KEK list of the minimum nodes that can include all user nodes included in G_i. For example, when $G_{data1} = \{u_1, u_3, u_4, u_7, u_8\}$, $KEK(G_{data1}) = \{KEK_1, KEK_{34}, KEK_{78}\}$ is established. The group update procedure is as follows.

The metadata server applies a data deduplication process with the user, and issues ownership of data M_i to user u_2.

The metadata server includes the user in ownership group G_i of data M_i. Then, $KEK(G_i)$ is updated. The added KEK is assumed to be KEK_2.

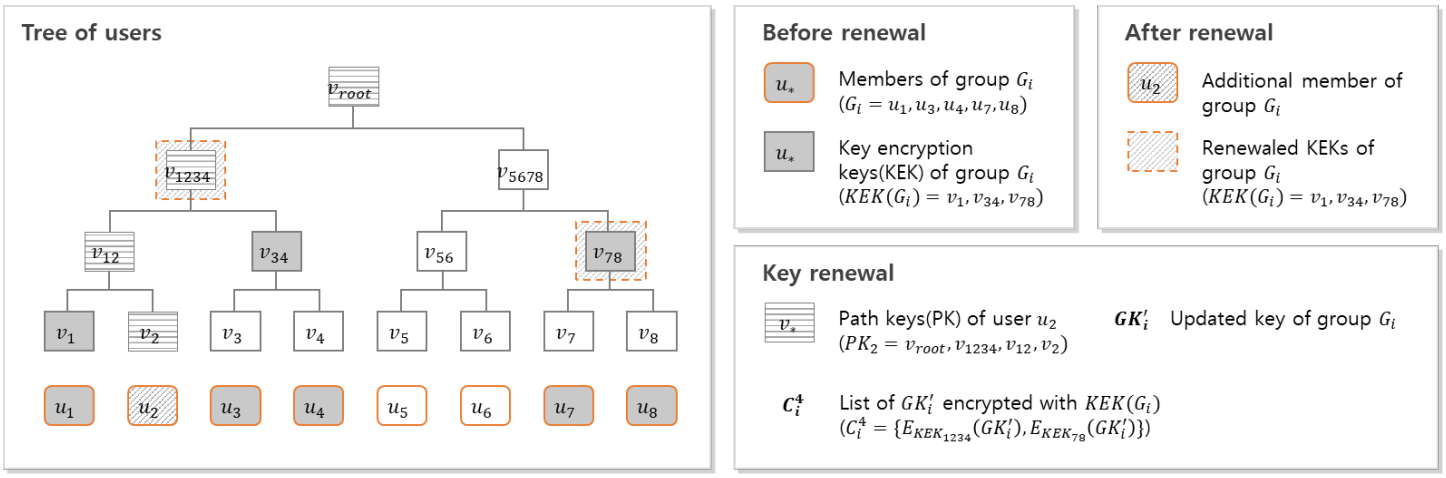


Figure 16. Method of renew ownership

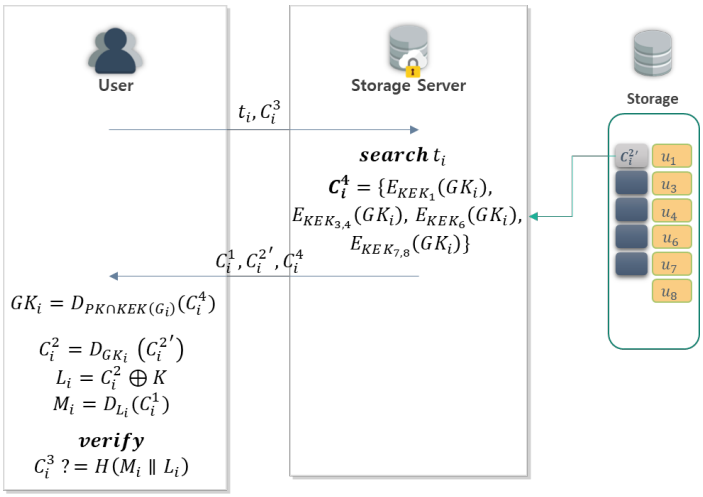


Figure 17. Data download phase

KEK(G_i) is composed of KEK, which can cover all users included in group G_i with the minimum KEK by combining the added and existing KEK.

The metadata server encrypts C_i^4 with GK_i, and encrypts GK_i with KEK_{1234}.

4.6. Data Download Phase

This procedure corresponds to Fig. 17, and is conducted when a user who has been issued ownership of data $M_i$requests data to be downloaded from the server.

The user transmits identifier t_i of the data to be downloaded and verification data C_i^3 to the metadata server.

The server searches for data corresponding to t_i, and then transmits $C_i^1, C_i^{2'}$, and C_i^4 stored in the server to the user.

The user obtains M_i using $C_i^1, C_i^{2'}$, C_i^4, and the user's own PK_u and k_i.

$$GK_i \leftarrow D_{PK \cap KEK(G_i)}(C_i^4) \tag{13}$$

$$C_i^2 \leftarrow D_{GK_i}(C_i^{2'}) \tag{14}$$

$$M_i \leftarrow D_{L_i}(C_i^1) \tag{16}$$

$$L_i \leftarrow C_i^2 \oplus K \tag{15}$$

The user compares C_i^3 and $C_i^{3'}$. This allows verification of the data integrity.

$$C_i^3 ?= H(M_i \| L_i) \tag{17}$$

5. Analysis of Proposed Scheme

In this paper, we propose a more efficient and secure method for deduplication of encrypted data based on Hur et al.'s dynamic ownership idea [9]. Therefore, it has similar features to the Hur et al. method in certain areas, but differs in the following ways.

Confidentiality: Cloud storage is continuously exposed to data breaches. Therefore, data stored in the cloud must be encrypted and stored. However, because data deduplication and data encryption have opposite characteristics, we used CE to apply them concurrently. CE is a technique for obtaining an encryption key K by hashing the data source M as in (18). Therefore, the same encryption key and ciphertext are always generated for users who own the same data source. Therefore, using CE, it is possible to deduplicate the encrypted data. In this proposed scheme, data encryption is conducted using RCE mode of MLE based on CE.

$$K \leftarrow H(M) \tag{18}$$

- **Integrity:** The user who has downloaded the data can verify whether the downloaded data have been transformed through the RCE tag validation. With this process, the user obtains data $chunk_i$ and conducts a hash operation using the encryption key L_i to create C_i^3, as shown in (19), which can be verified.

$$C_i^3 ?= H(M_i \| L_i), \tag{19}$$

- **Anonymity :** This method is based on Hur et al.'s approach. When a user requests a data download, the cloud storage server sends the encrypted group key to the user. A user with legitimate ownership can decrypt the key and can verify that

Table. 1. Comparison with other technologies

				Hur, et al.	Kim, et al.	Proposed Scheme
Deduplication Location				Server Side	Client Side	Client Side
Poison Attack Resistance				△	○	○
Proof of ownership				○	X	○
Dynamic Ownership Management				○	X	○
User Anonymity				○	X	○
Compu-tation	Upload	First upload	User	2H + 1SE + 1⊕	3H + 1SE	3H + 1SE + 1⊕
		Subsequent upload	User	2H + 1SE + 1⊕	3H + 1SE	3H + 1⊕
		Poison attack occurred	User	2H + 1SE + 1⊕	3H + 1SE	4H + 1SE + 1⊕
	Download		User	1H + 3SE + 1⊕	1H + 1SE	1H + 3SE + 1⊕
	Ownership group renewal		Server	1H + 3SE + 1⊕	1H + 1SE	1H + 3SE + 1⊕

○: Offer; X: Not offer; △: limited offer;
H: Hash algorithm; SE: Symmetric key encryption; ⊕ : XOR Operation;

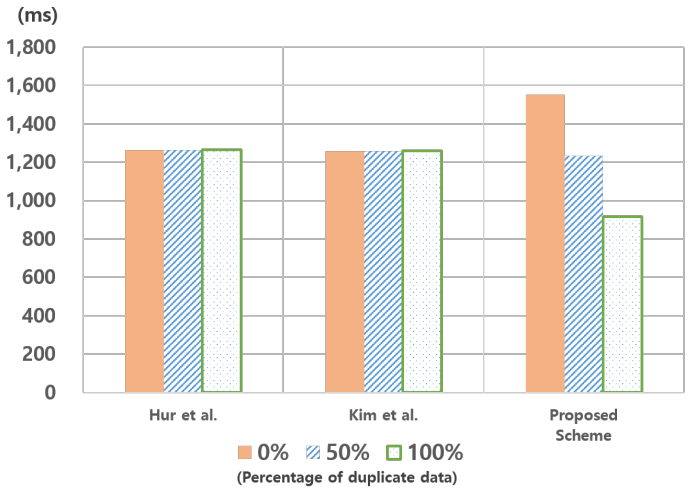


Figure 18. Computation time of 250k blocks

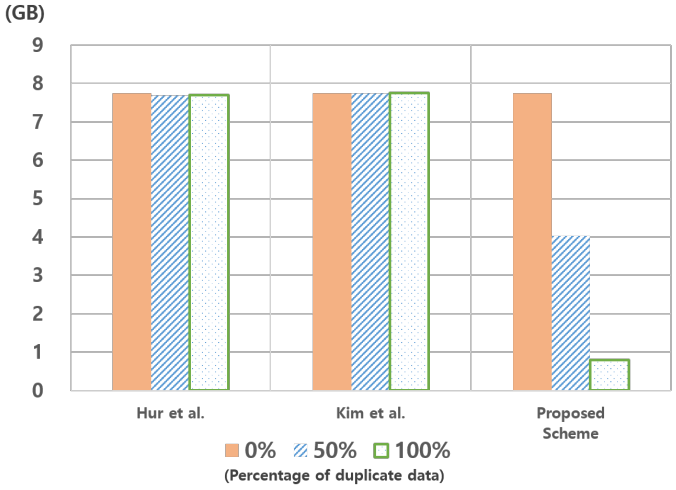


Figure 19. Data traffic of during deduplication of 250k blocks

it has legitimate ownership without identifying the user. Accordingly, the user has ownership of the corresponding data through this method, and thus downloads and decrypts the data through the process shown in (13-16).

- **Efficiency:** The proposed method shows that the total computational load decreases as the duplication ratio increases, the cause of which can be observed in Table 1. With the proposed method, one more hash operation per block is generated than in Hur et al.'s and Kim et al's method during the upload process when the first upload and poison attack occur[11, 9]. However, in the proposed scheme, the 3H + 1SE operation is omitted when redundant data exist. Therefore, the total number of operations can be reduced as the ratio of redundant data increases. In addition, the proposed scheme uses client-side deduplication. This method can conduct data uploads and deduplication with less computational and transmission overhead than server-side deduplication methods when uploading duplicated data, as shown in Fig 18, 19[10].

- **Resistance to poison attack:** In Hur et al.'s scheme, when a poison attack occurs, its occurrence can be checked at the download phase. However, data lost owing to a poison attack cannot be recovered. Therefore, the proposed scheme detects the poison attack in the subsequent upload phase to prevent data loss. In Hur et al.'s scheme, when a poison attack occurs, its occurrence can be checked during the download phase. However, data lost from a poison attack cannot be recovered. Therefore, the proposed scheme detects a poison attack during the subsequent upload phase to prevent a data loss [8].

- **Resistance to dictionary attack:** CE is vulnerable to dictionary attacks. Basically, because CE has a structure for obtaining an encryption key from a data source, the encryption key can also be obtained if the data source can be guessed. Therefore, to achieve resistance to a dictionary attack, data encryption key acquisition through data guessing should be made impossible. For this purpose, the proposed scheme conducts data encryption using RCE mode of MLE based on

CE. In RCE mode, the key obtained by hashing the data source is not used as a data encryption key, but is used as a key "K" for encrypting the data encryption key "L." The formula for this is shown in (8),(21–23). Therefore, even if a data source is guessed, it is possible to resist a dictionary attack, which is an attack achieved through data guessing, because it does not obtain the data encryption key directly, but acquires a key capable of decrypting the data encryption key.

$$L \xleftarrow{\$} \{0,1\}^{\lambda(K)} \tag{21}$$

$$C^1 \leftarrow E_L(M) \tag{22}$$

$$C^2 \leftarrow L \oplus K \tag{23}$$

6. Conclusion

This study was conducted to improve the efficiency and security of data deduplication. Cloud storage wastes space owing to the redundant storage of the same data. Data deduplication has been proposed to solve this problem. Data deduplication is a technology that saves storage space by preventing the same data from being stored. Therefore, the uploaded data are compared with the data stored in the existing storage to make sure the data are the same. However, cloud storage always incurs a threat of data leakage, and thus data encryption should not be used to identify the data source. Therefore, cloud storage requires the data to be encrypted and archived through deduplication. However, CE was proposed to deal with the conflicting characteristics of data encryption and deduplication technologies. Because CE is a technology for generating an encryption key by hashing the data source, the same ciphertext is always generated when encrypting the same data source. It therefore becomes possible to deduplicate the encrypted data. However, CE poses a threat of a dictionary attack, and client-side deduplication poses a threat to a poison attack. Therefore, we use RCE mode of MLE to solve this problem. MLE is a CE-based encryption technology, and can prevent data-source guessing. In addition, the verification data generated in RCE mode can be used to identify the occurrence of a poison attack. In addition, when data deduplication is conducted, the ownership information of the user is updated. Because cloud storage is an environment used by multiple users, such ownership information is frequently updated. Therefore, ownership management difficulties may occur, and the identity of the user may be exposed when using such ownership information. To solve this problem, we applied ownership management technology using a binary tree based on Hur et al.'s scheme. Through this, this paper is safe from poison attack, dictionary attack, proof of ownership, ownership management and ownership anonymity. In addition, by applying a client-side deduplication environment, the number of computations and amount of communication can be reduced as the data redundancy ratio increases. As a result, the proposed scheme is effective against various security threats, and reduces the amount of computational traffic as the data redundancy ratio increases.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2016R1D1A1B03935917) and Barun ICT Research Center at Yonsei University.

References

[1] Kim, K. W., Joo, Y. H., Eom, Y. I. "Technical trends for cloud storage data deduplication", Proceedings of Symposium of the Korean Institute of Communications and Information Sciences, 2012, pp. 228–229. DOI: 10.22648/ETRI.2018.J.330107

[2] Kim, W. B., Lee, I. Y., Ryou, J. C. "Improving dynamic ownership scheme for data deduplication." Computer Applications and Information Processing Technology (CAIPT), 2017 4th International Conference on. IEEE, 2017. DOI: 10.1109/CAIPT.2017.8320671

[3] Storer, M. W., Greenan, K., Long, D. D., Miller, E. L "Secure data deduplication", Proceedings of the 4th ACM International Workshop on Storage Security and Survivability, ACM, 2008, pp.1–10. DOI: 10.1145/1456469.1456471

[4] Douceur, J. R., Adya, A., Bolosky, W. J., Simon, P., Theimer, M. "Reclaiming space from duplicate files in a serverless distributed file system." Distributed Computing Systems, 2002. Proceedings. 22nd International Conference on. IEEE, 2002. pp. 617-624. DOI: 10.1109/ICDCS.2002.1022312

[5] Bellare, M., Keelveedhi, S., Ristenpart, T. "DupLESS: Server-Aided Encryption for Deduplicated Storage." IACR Cryptology ePrint Archive 2013 (2013): 429.

[6] Kaaniche, N., Laurent, M. "A secure client side deduplication scheme in cloud storage environments." NTMS 2014: 6th International Conference on New Technologies, Mobility and Security. 2014. pp. 1-7. DOI: 10.1109/NTMS.2014.6814002

[7] Bellare, M., Keelveedhi, S., Ristenpart, T. "Message-locked encryption and secure deduplication." Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer Berlin Heidelberg, 2013, pp. 296-312. DOI: 10.1007/978-3-642-38348-9_18

[8] Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, A. "Proofs of ownership in remote storage systems", Proceedings of the 18th ACM conference on Computer and communications security. ACM, 2011, pp. 491-500. DOI: 10.1145/2046707.2046765

[9] Hur, J., Koo, D., Shin, Y., Kang, K. "Secure Data Deduplication with Dynamic Ownership Management in Cloud Storage", IEEE Transactions on Knowledge and Data Engineering, 28(11), 2016, pp. 3113-3125. DOI: 10.1109/TKDE.2016.2580139

[10] Kim, K., Youn, T. Y., Jho, N. S., Chang, K. Y. "Client-Side Deduplication to Enhance Security and Reduce Communication Costs." ETRI Journal 39(1), 2017, pp. 116-123. DOI: 10.4218/etrij.17.0116.0039

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 270-275 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Sequential Card Activities of Recollect Experience for Elderly with Dementia

Fang Lin Chao*, Boxiu Fanjiang

Department of Industrial Design Chaoyang University of Technology, 436, Taiwan R.O.C

ARTICLE INFO

Article history:
Received: 11 May, 2018
Accepted: 28 July, 2018
Online: 10 August, 2018

Keywords:
Sequential card
Dementia
QR code
App.

ABSTRACT

In this study, the authors propose a modified design of sequential card activities to recollect experience for the elderly with Dementia. The cards and App were designed to complement each other for Nostalgia-based support. Image and text are utilized to trigger memories that bring out shared experience from the patient. By using suitable trigger elements, group sharing, and physical objects, one can enhance the memory recall. The procedure of rice dumpling preparation and making were listed in a sequential step and printed on cards. Cards are shuffled, and the elderly are asked to sort these cards to right sequence with the help of QR scanner. In multiple theme scenarios, different groups of object making steps were mixed, and the player needs to select cards that related to a specific group. Using actual items and activities, the triggered of elder's experience was confirmed. The qualitative testing results showed that the elderly enjoyed the activities. Those events verified the effectiveness of the proposed method.

1. Introduction

There had been an increasing trend for the number of patients with dementia in recent years, as well as a trend in which the average age for patients with dementia is becoming younger. This situation has been a challenge in the country where the amount of caregivers is low compared to the elderly population. The medical community has not found a cure for dementia; however, the condition often follows some pattern. Characteristically, foreign workers hired as caregivers do not make use of community-based services as they care for the elderly. Many efforts had been developed to support the elderly with dementia. For instance, training program, caring direction and technologies have been utilized to investigate and improve patient well-being [1-3]. The designer should realize that an integrated approach is necessary to address the cognitive, emotional and motivational aspects.

Although low-tech external memory aids have limitations, these limitations may be addressed by smart assistive technologies. Mihailidis and Fernie had concluded in a study that people with dementia assisted by context-aware computing [4]. "The research illustrates developmental training needs for dementia nursing to improve the efficacy of assessments and therefore interventions" was proposed by Carradice et al. [5]. Banerjee discuss the quality of care in dementia, co-resident spouse care with Alzheimer's dementia. Training benefits on cognitive, behavioral, and global functioning in patients with Alzheimer's disease were reported [6]. If possible, technologies should be designed to be friendly and noninvasive.

We had observed that some patients fail to remember recent events quickly, but can recall events from the distant past. To address this, designers need to include an individual's environmental context and social relationship during their design process. Group-based nostalgia induced recollect experience may take different forms in the lives of each elderly. "During bus trip in a distant city, nostalgia was identified as a key element of the experience for socializing" [7]. In elder care center in Taiwan, some may reminisce through sing old traditional songs together, while others may look at old photos of buildings or sceneries, or read letters from the past from acquaintances. Recollect experience can be an image, text or sound track that can trigger the memory. In the instance where the caregiver and the elder have no close relationship, the activity seldom triggers any feedback. This problem indicates that the event is not sufficient, and the reaction was rendered as non-existent. As shown in Figure 1, it's difficult to communicate since they may not know much about the elder's background, so professional activities are needed.

2. Objectives and Methods

The goals and approaching of this study are:

1. Design sequential card activities of recollect experience that are suitable for patients with memory loss; trigger feedbacks from them using the appropriate images and text.

*Corresponding Author: Fang Lin CHAO, email: flin@cyut.edu.tw



Figure 1. Relationships between person help recollect experience: A foreign helper has less in common with the elder, which causes less memory recollect .

2. To propose a group activity, through the mutual agitation among members; the seniors with dementia can discover the differences between others and arouse interest. Based on personal characteristics, actual items were utilized to trigger past experience.

An improved design was used so that participants can use it actively without technology barrier. In the nostalgia-based supporting, different presentation of information and background memory elements can enrich interaction. During this study, the designer and the elderly play with the sequential card game together. The designer can test the system, collect feedback, and implement improvements in the newer version. Sections III, IV and VI describe our evolutional development cycles from the initial concept.

3. Arrange Sequential Card Activities

There are memories or facts in memory backtracking. Consciousness not only contains symbols, words, or phenomena but a combination of life experiences. The feature of cards game is that each card has its facts and events. When any card wakes up the user's memory, he can continue to explore the relationship between the cards. At the time of exploration, the life experience began to be excavated. The card game proposed in the previous reference applied to the wake of the experience of the elderly [8]. This article will further elaborate on the relevant content and details.

3.1. Design approach

Information technology used in many ways. For example, we can implement card games within touchscreens with software; because the one operating the app needs to be familiar with the use of mobile phones, the elderly may not suitable. Although the design has the greatest flexibility, we still adopt "supportive" proposals. The APP is only a supporting character. If there is any doubt, one can use a scanner to get help. It is as though we are asking others a question. Because the sequential card games are with a physical object that can be touched, it is suitable to link with the life experience of the elderly and reduces the fear.

3.2. Concept realization

Software is used as a means to reduce the stress in taking care of elderly with dementia; since they often repetitively ask the same questions, the average caretaker may be irritated from answering the same questions over and over again. Thus, the app is advantageous in that it can make a reply anytime, and answer related questions immediately, and won't be irritated from repeating the same answers.

Table 1. The concept shifts concerning the user confidence, where fluency means "the possibility of without obstacles during the game."

	standalone within screen	supportive explanation
flexible	+++	- -
fluency	+	++
affordance	- - -	+++

Based on Table 1, we utilize app as a supporter or explainer. Sequential card was designed (Figure 2) with the theme of wrapping rice dumplings (zongzi). For each step card, a proper image and description text should be matched. The cards are shuffled in random order before the game, and the elderly are asked to sort these cards into the correct sequence.

As technology is added in design, we have to consider whether the senior will reject or not feel accommodated to using these tools. Thus, in the card design, pictures or text is still shown on the front side, while the tag placed on the back side. The user won't often be distracted by tag. When the user wishes to seek help, they can get pertinent information by moving the card to a fixed position. QR code was placed on the back of each card. While the elderly see the image, the scanner read the QR code and activate necessary audio explanation with adjustable volume and speed (Figure 3, 4). During the process, if the elderly ever want to express anything about personal experience, question or feedback, the elder's voice response would be collected.

Remaking the step cards would be inconvenient when there are plenty of triggering elements. A reconfigurable card holder can be used to different sequential cards. Considering game's flexibility, while selecting the proper subject, the card's corresponding QR code [8] will not require re-setting.



Figure 2. Continuous applicable card holder. A reconfigurable slot is placed on the side wall for inserting a card. A user can change the theme more efficiently with this design.



Figure 3. Scanning will take some time for elderly.



Figure 4. QR code placed at the back of the sequential card. With an unstable hand, the user need moves the card to the designated position stably to scan the code.

During the trial, we noticed a few unstable hands during code scanning. In the modified design, QR code is placed at the back of the card to make it more convenient for the elderly. When the user can't hold the phone steadily with their hands, scanning will take some time. We can make the response time from scanning faster by placing the phone in a fixed position (Figure 5). If the user can hold the phone steadily and successfully operate the app, then a fixture for scanning won't be necessary.



Figure 5. When the user can't hold the phone steadily with their hands, the supporting fixture make the response time from scanning faster

4. Mobile App Design

The app was made using MIT App Inventor 2 which use module block in programming. App design arranged procedure screens as Figure 6. App Inventor programs blocks (Figure 7) represent sensing the measured value through the QR connection.



Figure 6. When the user clicks on an image, a scanner pops up. As the user scans the corresponding image on the card, texts detailing each step of the food preparation process pops up.

Students in the industrial design department seldom use programs, so they often forget the syntax and structure of the program. The module block can connected to another neighborhood block. It will be pushed apart automatically if it is not appropriate. Therefore, the designers pay attention to the main logical relationship between the program block and the selection of the module. The connection between Bluetooth blocks, we need to setup the paired connection and send the appropriate data to the other party. As shown in Figure 7, the program block and the link of the modules is quite simple; although the blocks are simple, they still achieve powerful functions.

An event is an action that takes place. App Inventor II uses event handler blocks to specify how the program responds to individual circumstances. There is a user-initiated and automatic event. Every time the interval is up, a timer event triggered. Build-in sensors also initiated events when the user made a specific movement or orientation.

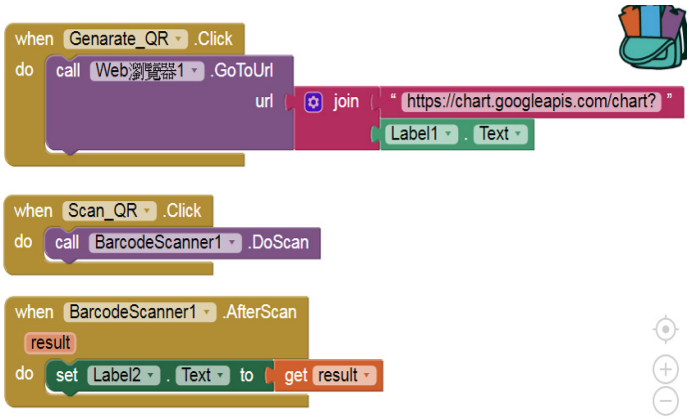


Figure 7. QR scan programming with modulated App Inventor blocks.

5. Multiple theme Scenarios

The more advanced games are organized for higher-skilled people; three recollect experience activities are available with different series of cards. Six step card blocks are cut to be in a 45-degree plane. By touching the magnets on each side, we can assemble six side blocks into a cube (Figure 8).

In multiple theme scenarios, different object making process are mixed, the player needs to select cards that related to specific objectives. After grouping those cards, one need to arrange those cards in proper sequence. When the blocks are organized into a cube, there is more variation for the player. The player cannot see all figures at once. The elderly need rotate the cube, and think about the category relationship between the images, and try to exercise their memorization ability more.

Assuming three recollect experience activities are available (making of rice dumpling, red turtle cake, and rice noodle soup), and each of these three groups has 6 step cards to be reordered. Put together cards of the same order, and make six cubes, and give those to the user. The researcher provides a questionnaire to the player. For example, if a user chose a turtle cake theme, the user has to make the related sides of each cube face upwards and rearrange these cubes.



(a)



(b)

Figure 8. Three-dimensional sequential card game, (a) magnets on the edge of a card allows each card to be attached to the cube, (b) the combined cube. In multiple theme scenarios, each card belongs to a different theme.

The flow chart for the sequential card game indicated in Figure 9. The explanations of each card are stored at separate lists. Player need recall experience about the steps of trigger elements and the timing relationship between cards. The cards help people focus and trigger the memory of a player.

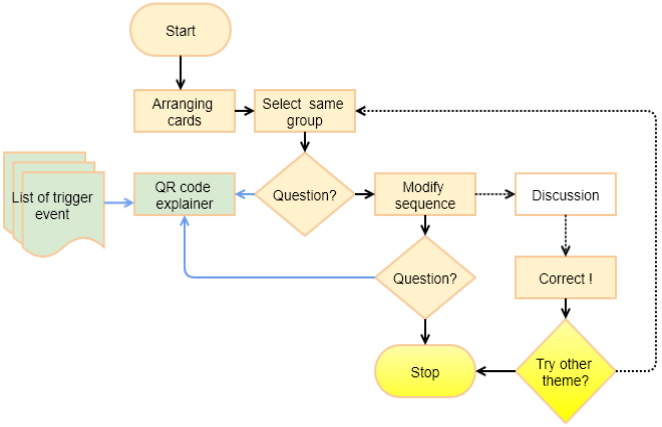


Figure 9. Sequential card and app flow chart indicates recollect experience with the support of QR code explainer. Question and discussion may be triggered while modifying sequence. After finish the sequential placement, users can try another theme.

6. User Testing

6.1. Elderly center interview

During the evaluation process, the participants actively talked to us about their experiences in making rice dumplings. For example, one mentioned a tip on how to make tastier dumplings, and one made a suggestion based on how her mother made it. The two testers shared the experiences with each other, which enhanced the recollect experience treatment's effect. The lady (in Figure 10) talked about when she was little, her family made red turtle cake during Christmas time, because her family treats Christmas like Lunar New Year, and this was their way of celebration. Tester C came from mainland China. She is a retired teacher. She is quiet and not very talkative to others. She likes to cook at home; she usually eats noodles and knows how to wrap rice dumplings herself. She prefers plain or slightly sweet rice dumplings with red bean paste or red dates added. Tester D is also a quiet lady with dementia. She often repeats the same thing over and over. She has more of a monotonous lifestyle and regularly watches TV. She likes to cook at home; usually, eat rice and prefer meat dumplings.

During the evaluation process, participants had made feedback about the using the sequential card. The shared message was recorded by a researcher. A five points Likert scale survey about playing the game were collected, the averaged of the score are shown in Table 2. The single theme achieved better manipulation experience, less skill and explanations were requested. When four people play a single theme sequential card game together, the games are fluent and emotional engagement (talkative to others) are high (score above the average value). This activities indicated positive feedback. They are grateful to share individual experiences. The communication links those elderlies together, and make nostalgia-based treatment more effective.

The multiple theme achieved less manipulation experience, more skill and memory function are request for proceed the game. More explanations were requested. Arranging magnets placed at sides and organized into a cube is a challenge for elderly with less detail manipulation skill. We suggested multiple theme is suitable only for advanced users.



Figure 10. User observation of multiple theme, to provide proper image with good resolution and teach user to get explanations through the support of QR code.

6.2. Follow-up personal activity

Through group activities, the mutual interaction among members, old people with dementia discovered the differences of others. Then, a follow-up private action based on actual items and

activities was conducted to confirm when measurable memory is triggered.

Table 2. averaged feedback about using the product (5 points Likert scale)

	single theme	multiple themes
manipulation skill	4.25	3.25
request explanation	3.00	4.50
fluent of game	4.25	3.00
emotional engage	4.25	3.25
talkative to others	3.75	3.00

Our participant is a 90-year-old woman with dementia; she often repeats the same thing and does not notice it. She has a monotonous lifestyle, and usually watches TV, and sleeps for longer durations. Based on past life experience, the selected personal activity is rice dumpling making. We prepared objects such as bamboo leaves, strings, and red dates (Figure 11).



Figure 11. The prepared objects and items during personal practice and interview.

Through the initial guidance, the participant gradually recall her past experiences with rice dumpling making. We then made recordings of her descriptions with different aspect of experiences (Figure 12).

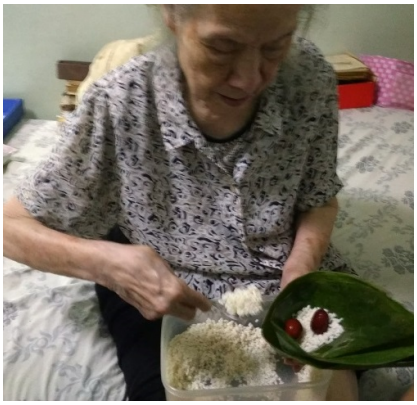


Figure 12. Personal practice and interview

Second, we will analyze the sentences (underline) in her descriptions, and categorize them into facts, estimations based on events, phenomena that may result from situations, and personal emotions. The subject's narrative was classified, and her depth of participation is presented in Table 3. During 3 hours of communication, we collect 24 minutes of voice records which are summarized below. Some paragraph was repeated to show her traits.

1. *How many kilograms of rice do you have to buy?* Now you can only get one serving. You have to buy a kilo of rice and soak the rice in water. Bamboo leaves must be soaked before they become soft.
2. *Do you know how to wrap a rice dumpling?* I thought these [strings] were worms! We put red beans with sugar, like cooking porridge. Rice must wash first. A cup of rice is fine. The rice must wash for preparing quickly.
3. Bamboo leaves must first be soaked water to be softened, which is not easy to be torn. Just put a few red dates on it. Wash them first and soak them in water, and throw a few into the rice dumplings. It will swell as soon as it cooked, and it will be delicious.
4. *What is the use of dates?* Red dates are used for eating ah! The boiled berry swell while cooking. There are hard shells inside the red dates, and your teeth may hurt if you accidentally bite on them. The price of bamboo leaves is high. Someone in the market sells USD, and another sells only $2. Larger leaves are more expensive because fewer people are selling them.
5. *When there is no bamboo leaf, how should you pack the rice dumplings?* Smaller leaves can be used, a few pieces together. If you don't break it, you can dry it and re-use it. Bamboo leaves have a scent! Maybe someone thinks it's smelly odor.
6. *How do you buy the leaf?* I guess. Do you mean to purchase bamboo leaves? Would you put pork or salted eggs in dumplings? We usually eat sweet rice dumplings. Let's start wrapping! I dumped the water and, here is the material for you, bamboo leaves and rice, and strings.
7. *Why is water dripping out from the rice dumplings?* There is water flowing out, and a hole on the side of the rice dumplings and rice inside may come out. Why do you use two leaves to wrap it? Because a piece of a leaf will quickly be torn with force and with two pieces, it will be firm.
8. *How do you tie the rope while bundling?* Hold it with one hand and use the other to help. This one was too flat and not wrapped well. Someone will use teeth to help bite the string and tie it so that it's tighter.

Owing to the capacity disorder of poor hearing, the participant would often say "what you are saying I can't hear". Sometimes she asks, "Where do you come from," "Where do you go to work?" and similar typical questions. This is the typical response of dementia patients.

The details response of those activities contains logical relations and inferential information, which are different from her usual behavior. In addition to the characteristics of the sentence analysis of her recorded response, there is also the presentation of actions. For instance, each of the steps of wrapping a rice dumpling is hidden as procedural memory in the cerebellum. When the hand touches these objects, such as materials for wrapping a rice dumpling, those memories are evoked; and the previous situations can emerge.

Table 3. The summary of voice recording and categorizing into facts, estimations, inductions and emotions.

facts	estimation	induction	emotion
hard shells inside fewer people are selling soak the rice red dates with sugar, wash first usually eat sweet water flowing out	hard shells inside only get one serving. buy a kilo of rice a cup of rice rice inside may run out will be firm two leaves to wrap	become soft will swell will be delicious cooked quickly Small pieces dry it and re-use it use teeth to help have a scent	I thought these were worms! Red dates are used for eating ah! I guess. Do you mean to buy bamboo leaves someone thinks it's smelly odor

The academic reference [9] indicates “Imagery and motion were found to share common cerebellar circuitry;” “which reflect a higher demand for the cerebellum as compared to a motion." The activation pattern observed during the execution of the finger-to-thumb movement (motor preparation and timing) is in agreement with theories [9]. “Specific localizations of right superior cerebellar activation, the reaction times were significantly increased for the verbal working memory and motor control tasks” were reported [10]. From the academic reference and user observation, the presentation of actions is another crucial trigger element. It may conduct in the future design of sequential card game.

7. Conclusion

The main contributions in the paper are an implementation of the sequential card activities with different themes to recollect experience. The steps for rice dumpling preparation and making were printed on cards. The elderly are asked to sort these cards to right sequence with the help of a QR scanner. The qualitative testing results showed that the elderly enjoyed the activities. Through the game and trigger elements, non-family members or caregivers have more social interaction and involvement in personal experiences. The communication appeared among members during the group activity. The old people with dementia discovered their differences and enabled passing experiences between each other.

Based on individual characteristics, we also conducted actual items and physical activities; the participant is triggered to find out his experience that has been identified. Concerning unique features of dementia, understanding of the activation made by the elderly was confirmed. In future, the app will extend with remote communication ability to share with the friends. The presentation of motion actions is another crucial trigger element. It will conduct in future card game design.

References

[1] M.S. Bourgeois, C. Camp, M. Rose, B. White, M. Malone, J. Carr, and M. Rovine, “A comparison of training strategies to enhance use of external aids by persons with dementia,” J. Commun. Disorders, **36**, 361-378, 2003. https://doi.org/10.1016/S0021-9924(03)00051-0

[2] L. Clare, B.A. Wilson, G. Carter, and J.R. Hodges, "Cognitive rehabilitation as a component of early intervention in Alzheimer's disease: A single case study," Aging Mental Health, **7**, 15–21, 2003. https://doi.org/10.1080/1360786021000045854

[3] S. Josefsson, L. Backman, L. Borrel, B. Bernspang, L. Nygard, and L. Ronnberg, "Supporting everyday activities in dementia: an intervention study,” Int. J. Geriatric Psychiatry, **8**, 395–400, 1993. https://doi.org/10.1002/gps.930080505

[4] A. Mihaildis and G. Fernie, “Context-aware assistive devices for older adults with dementia,” Gerontechnology, **2**, 173–189, 2002. https://doi=10.1.1.588.447&rep=rep1&type=pdf

[5] Carradice, Angela, Marie Claire Shankland, and Nigel Beail. "A qualitative study of the theoretical models used by UK mental health nurses to guide their assessments with family caregivers of people with dementia." International Journal of Nursing Studies 39.1 (2002): 17-26. https://doi.org/10.1016/S0020-7489(01)00008-6

[6] Banerjee, Sube, et al. "Improving the quality of care for mild to moderate dementia: an evaluation of the Croydon Memory Service Model." International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences 22.8 (2007): 782-788.

[7] S. Fairley “In search of relived social experience: group-based nostalgia sport tourism,” Journal of Sport Management, **17**(3), 284-304, 2003. https://doi.org/10.1123/jsm.17.3.284

[8] Chao, Fang-Lin, et al. "Design Jigsaw puzzle and app for Recollect experience support on elderly with Dementia." Awareness Science and Technology (iCAST), 2017 IEEE 8th International Conference on. IEEE, 2017. DOI: 10.1109/ICAwST.2017.8256464

[9] Luft, A.R., Skalej, M., Stefanou, A., Klose, U. and Voigt, K., 1998. Comparing motion‐related and imagery‐related activation in the human cerebellum: A functional MRI study. Human brain mapping, 6(2), pp.105-113. https://doi.org/10.1002/(SICI)1097-0193(1998)6:2<105::AID-HBM3>3.0.CO;2-7.

[10] Desmond, John E., SH Annabel Chen, and Perry B. Shieh. "Cerebellar transcranial magnetic stimulation impairs verbal working memory." Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society 58, no. 4 (2005): 553-560. https://doi.org/10.1002/ana.20604

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 271-282 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Steganography System with Application to Crypto-Currency Cold Storage and Secure Transfer

Michael J. Pelosi[*,1], Nimesh Poudel, Pratap Lamichhane[1], Danyal Badar Soomro[2]

[1]*Computer Science and Technology Faculty, East Central University, 74820, USA*

[2]*School of Software Engineering, Chongqing University, Chongqing 400044, China*

A R T I C L E I N F O

Article history:
Received: 15 November, 2017
Accepted: 05 February, 2018
Online: 25 April, 2018

Keywords:
Steganography
One-time pad
Steganalysis
Information hiding
Digital forensics

A B S T R A C T

In this paper, we introduce and describe a novel approach to adaptive image steganography which is combined with One-Time Pad encryption and demonstrate the software which implements this methodology. Testing using the state-of-the-art steganalysis software tool StegExpose concludes the image hiding is reliably secure and undetectable using reasonably-sized message payloads (≤25% message bits per image pixel; bpp). Payload image file format outputs from the software include PNG, BMP, JP2, JXR, J2K, TIFF, and WEBP. A variety of file output formats is empirically important as most steganalysis programs will only accept PNG, BMP, and possibly JPG, as the file inputs. In this extended reprint, we introduce additional application and discussion regarding cold storage of crypto-currency account and password information, as well as applications for secure transfer in hostile or insecure network circumstances.

1. Introduction

In this paper, we introduce a comprehensive steganography software system and platform framework based on One-Time Pad (OTP) encryption and adaptive steganography technology. We provide usage recommendations and advice guidelines. The system is tested and shown to be resistant to many common steganalysis attacks. In the context of this paper we are assumed advocate of the steganography; someone who may be a political dissident in an oppressive regime, a religiously persecuted individual, a friendly agent engaging in covert communication, or a lawful individual desiring complete communication privacy, among other compelling examples.

2. What is Steganography?

Steganography, the art of invisible communication, is achieved by hiding secret data inside a carrier file such as an image. After hiding the secret data, the carrier file should appear unsuspicious so that the very existence of the embedded data is concealed. A major drawback to encryption is that the existence of the message data is not hidden. Data that has been encrypted, although unreadable, still exists as a suspicion arousing file transfer. If given enough time, once alerted, someone could potentially decrypt the data or derive other intelligence regarding either sender or receiver. A solution to this problem is steganography. This is the ancient art of hiding messages so that they are not detectable.

In steganography, the possible cover carriers are unsuspicious appearing files (images, audio, video, text, or some other digitally representative code) which will hold the hidden information. A message is the information hidden and may be plaintext, cipher text, images, or anything that can be embedded into a bit stream. Together the cover carrier and the embedded message create a stego-carrier. Hiding information may require a stego key which is additional secret information, such as a password or OTP key, required for embedding the information. For example, when a secret message is hidden within a cover image, the resulting product is a stego-image. A possible formula of the process may be represented as: cover medium + embedded message + stego key = stego-medium.

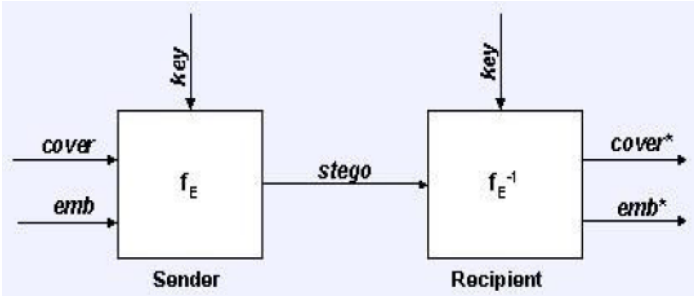


Fig:1 Graphical Version of Steganographic system

*Corresponding Author: Michael J. Pelosi, East Central University,
Email: mpelosi@ecok.edu

f_E: steganographic function "embedding".
f_E^{-1}: steganographic function "extracting".
cover: cover data in which emb will be hidden.
emb: message file to be hidden.
stego: cover data w

The advantage of steganography is that it can be used to secretly transmit messages without the fact of the transmission being discovered. Often, using encryption might identify the sender or receiver as someone with something to hide. It is believed that steganography was first practiced during the Golden Age in ith the hidden message. Greece. An ancient Greek record describes the practice of melting wax off wax tablets used for writing messages and then inscribing a message in the underlying wood. The wax was then reapplied to the wood, giving the appearance of a new, unused tablet. The resulting tablets could be conveniently transported without anyone suspecting the presence of a message beneath the wax.

2.1. LSB Steganography

The simplest and popular image steganographic method is the least significant bit (LSB) substitution. It embeds messages into cover image by replacing the least significant bits directly. The hiding capacity can be increased by using up to 4 least significant bits (one each for Red, Green, Blue, and Alpha color channels, respectively) in each pixel. It has a common weak point i.e. the sample value changes asymmetrically. When the LSB of cover medium sample value is equal to the message bit, no change is made. Otherwise the value 2n is changed to 2n+1 or 2n+1 is changed to 2n. There are many improvements and modifications that have been proposed to strengthen this technique, such as adaptive techniques that alter payload distribution based on image characteristics. If the message is first encrypted and then embedded, the security is enhanced.

2.2. One-Time Pad

The "one-time pad" encryption algorithm was invented in the early 1900's, and has since been proven as unbreakable. The one-time pad algorithm is derived from a previous cipher called Vernam Cipher, named after Gilbert Vernam. The Vernam Cipher was a cipher that combined a message with a key read from a paper tape or pad. The Vernam Cipher was not unbreakable until Joseph Mauborgne recognized that if the key was completely random the cryptanalytic difficultly would be equal to attempting every possible key (Kahn 1996). Even when trying every possible key, one would still have to review each attempt at decipherment to see if the proper key was used. The unbreakable aspect of the one-time pad comes from two assumptions: the key used is completely random and the key cannot be used more than once. The security of the one-time pad relies on keeping the key secret and using each key only once.

The one-time pad is typically implemented by using exclusive-or (XOR) addition to combine plaintext elements with key elements. An example of this is shown in Figure 2. The key used for encryption is also used for decryption. Applying the same key to the cipher text results in the output of the original plaintext.

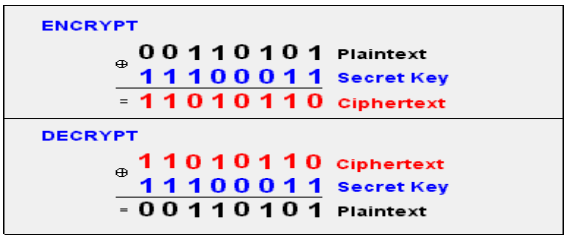


Fig.2 Examples of a One-time Pad implementation using XOR addition.

OTP is immune even to unlimited resources brute-force attacks. Trying all keys simply yields all possible plaintexts, all equally likely to be the actual plaintext. Even with known plaintext, such as part of the message being known, brute force attacks cannot be used, since an attacker is unable to gain any information about the parts of the key needed to decrypt the rest of the message.

3. Methodology

The following describes the general method implemented in the software for key generation, encryption, embedding, message transfer, and decryption.

- Random image keys are generated using a key generator program. The key generator program generates one-time pad keys that consist of random colored pixels. Each random colored pixel consists of random values for red, green, and blue colors throughout the image. One image key is generated for every message that is intended to be sent.
- To encrypt a message, a cover image and random key image is selected. Each pixel in the cover image is XOR'ed with the key image X, Y coordinate pixel. Each pixel consists of a 32-bit long integer color value. One byte each corresponds to red, green, and blue components, respectively. The XOR'ed pixel values are then adjusted to hide the message. The bytes in the message are divided up into bits — one bit per pixel. The least significant bit (LSB) in the XOR'ed pixel colors are then adjusted to hide the message. Bit values that do not correspond are adjusted (in general 50% of the values will already be set correctly). LSB's for red, green, or blue are selected based on a local pixel variation score, contingent if the sum of the RGB LSB's are even or odd (even corresponds to a 0 bit, odd to a 1 bit).
- At this point, the newly derived color values are XOR'ed once again with the random image key to generate color values very close to the original image. These pixel color values will be used to construct the steganographic image that will be sent to the receiver.
- Ideally at this point, both the original cover image and the senders copy of the random image key can be destroyed (forensically wiped from the hard drive using a file erasure procedure). This is to prevent later detection and statistical comparisons.
- Upon receipt of the steganographic image, the receiver loads the intended image key and XOR's each pixel of the steganographic image with its respective corresponding X, Y pixel in the image key. This will derive a series of bit values

that correspond to the plaintext message. The bits can be reassembled into bytes (and later 2-byte Unicode characters) that correspond to the plaintext message.

- The start and end of the message are delimited by randomly chosen 10-character delimiting strings that are embedded as EXIF comments into the random image key by the key generator program. Thus, random message padding is incorporated at the start and end of messages.
- The random image key also contains a random number seed, this is used for the random number generator algorithm in use and starts the generator at the proper sequence start value.

3.1. Random Number Generation

A cryptographically secure pseudo-random number generator (CSPRNG) or cryptographic pseudorandom number generator (CPRNG) is a pseudorandom number generator (PRNG) with properties that make it suitable for use in cryptography. Ideally, the generation of random numbers in CSPRNGs uses entropy obtained from a high-quality source, which might be a hardware random number generator or perhaps unpredictable system processes — though unexpected correlations have been found in several such ostensibly independent processes. Several robust CPRNGs are incorporated into the steganography software.

3.1.1 Mersenne Twister

The Mersenne Twister is a pseudorandom number generator (PRNG). It is by far the most widely used general-purpose PRNG. Its name derives from the fact that its period length is chosen to be a Mersenne prime. The Mersenne Twister was developed in 1997 by Makoto Matsumoto and Takuji Nishimura. It was designed specifically to rectify most of the flaws found in older PRNGs. It was the first PRNG to provide fast generation of high-quality pseudorandom integers. The most commonly used version of the Mersenne Twister algorithm is based on the Mersenne prime 219937−1. The standard library implementation of this, MT19937, uses a 32-bit word length. There is another implementation that uses a 64-bit word length, MT19937-64, that generates a different sequence. The software implements a cryptographically secure version of the Mersenne Twister provided by the algorithm authors Matsumoto and Nishimura.

3.1.2 Other Random Number Generators

Optional random number generator selections included in the OTP-Steg key generator program include the following (each of these can be optionally selected by the user):

- ISAAC— ISAAC (indirection, shift, accumulate, add, and count) is a cryptographically secure pseudorandom number generator and a stream cipher designed by Robert J. Jenkins, Jr. in 1996.
- CryptGenRandom-CryptGenRandom is a cryptographically secure pseudorandom number generator function that is included in Microsoft's Cryptographic Application Programming Interface. In Win32 programs, Microsoft recommends its use anywhere random number generation is needed.
- RtlGenRandom — On a default Windows XP and later install, CryptGenRandom calls into a function named ADVAPI32!RtlGenRandom, which does not require one to load all the CryptAPI classes for usage.
- Rnd () — Standard API random number generator (for research/testing purposes only – it is not cryptographically secure).

3.2 Key Delimiters

Upon key generation, a pair of key delimiters is also randomly chosen of 10 Unicode characters each for the start delimiter and end delimiter, respectively. These are used to indicate to the decryption program exactly where the message starts, and where it ends. Random padding is added to both ends of the message — the start and the end of the message embedded in the payload file. The key delimiters identify where to start the message text, and where to cut it short at the end of the message. These key delimiters are contained in the EXIF image comment data in the key file. No EXIF comment data whatsoever is contained in the payload file. Also, the key delimiter values are utilized for random number generation seed data used for encryption and decryption.

3.3 Expert System to Evaluate and Score Candidate Cover Images

It is well known from the literature that some cover images present much better candidates for steganographic security than others based on image characteristics. Typically, cover images with a high degree of pixel color variation, very few saturated white or black pixels, and few pixels next to each other of the same color, are excellent payload candidates. We implement an expert system to give the software user immediate knowledge of how good a candidate a potential color image is for detection security. We have incorporated a tentative scoring system that evaluates images based on several factors. The output score ranges from 0 to 100%, with greater than 90% score being a good candidate for a cover image. Scores of 80-90% are marginal, and less than 80% are considered not adequate. In the current preliminary version, a peak signal-to-noise ratio (PSNR) versus a solid color image is calculated. This rating is given a weighting of 25% in the overall score. Also, the number of same color pixels next to each other is given a weighting of 25% for up to 5% of the image pixels (in other words, a 5% of the image pixels are same color next to each other, this rating would be zero). Thirdly, a weighted rating of 25% is given to the number of white pixels, up to 5%. The same weighting is also calculated for black pixels. Each of the four factors is combined for the rating from 0% up to 100%. Ideally, a cover image will have zero white pixels, zero black pixels, very few colors next to each other that are the same, and a very high variation in color over comparison to a solid color image. Table 1 below lists the above and additional cover image scoring factors that could be evaluated in an expert system rating scheme.

3.4 Future Security = Small Payloads

To ensure robustness against potential future attacks we have limited payload relative sizes. The high limit for the bits-per-pixel pixel is approximately 25%. And since only half of pixels are typically altered based on the message, this corresponds to a practical limit of about 12.5%-pixel alteration. By limiting the

pixel bit payload, it quite robustly limits detectability now and in the future.

Table 1. Potential Candidate Image Scoring Factors.

Factor	Description	Value
PSNR over solid color	Peak signal-to-noise ratio of image to solid color image.	Higher values are better.
Percentage of saturated colors	Portion of the image that is either all-white or all black.	Lower values are better.
Percentage of nearby same colors	Portion of image that has neighboring pixels of the same color.	Lower values are better.
Randomness of LSB's	Measures of randomness of the distribution of the significant bits.	Higher randomness is better.
Random RGB LSB distribution	Randomness of each color channel.	Higher values are better.
RS test on Cover Image	Clean RS test on cover image.	Lower values are better — indicates less probability of a threshold being reached after encoding.
Chi-squared test on Cover Image	Clean Chi-square test on cover image.	Lower values are better.
Pure Photograph	Photo has not previously been compression encoded using algorithm like JPEG.	Straight from a high-quality digital camera is best.
Original Photograph	No other copies of the photo exist in clean or altered state that can be used for comparison.	Known source and originality is best here.
Dimensions	It is well known that extremely large images have less pixel color variation and steganography here is more easily detected.	Approximate pixel dimensions of images frequently found on the Internet are best — about 1600?200 or less pixels.

Extremely advanced statistical detection techniques are being promulgated that are improving the odds of successfully detecting steganography efforts. There is no guarantee that these steganalysis efforts will not double or triple in effectiveness in the next few years. As a safety measure and margin of security, payload size is strictly limited by the software to an amount that should be reasonably safe for the foreseeable future. This equals future security for messages that may be encrypted today and subsequently intercepted and archived for several years for later decipherment.

3.5 Steganalysis

Steganalysis is "the process of detecting steganography by looking at variances between bit patterns and statistical norms". It is the art of discovering and revealing covert messages. The goal of steganalysis is to identify suspected information streams, determine whether or not they have hidden messages encoded into them, and, if possible, recover the hidden information. Unlike cryptanalysis, where it is evident that intercepted encrypted data contains a message, steganalysis generally starts with several suspect information streams but uncertainty whether any of these contain hidden message. The steganalyst starts by reducing the set of suspect information streams to a subset of most likely altered information streams. This is usually done with statistical analysis using advanced statistics techniques.

Analyzing repetitive patterns may reveal the identification of a steganography tool or hidden information. To inspect these patterns an approach is to compare the original cover image with the stego image and note visible differences. This is called a known-carrier attack. By comparing numerous images, it is possible that patterns emerge as signatures to a steganography tool. Another visual clue to the presence of hidden information is padding or cropping of an image. With some steganographic tools if an image does not fit into a fixed size it is cropped or padded with black spaces. There may also be a difference in the file size between the stego-image and the cover image. Another indicator is a large increase or decrease in the number of unique colors, or colors in a palette which increase incrementally rather than randomly. These are just examples among the many published and effective approaches.

StegExpose is a steganalysis tool specialized in detecting LSB (least significant bit) steganography in lossless images such as PNG and BMP. It has a command line interface and is designed to analyze images in bulk while providing reporting capabilities and customization which is comprehensible for non-forensic experts. The StegExpose rating algorithm is derived from an intelligent and thoroughly tested combination of pre-existing pixel based steganalysis methods including Sample Pairs by Dumitrescu (2003), RS Analysis by Fridrich (2001), Chi-Square Attack by Westfeld (2000) and Primary Sets by Dumitrescu (2002). In addition to detecting the presence of steganography, StegExpose also features the quantitative steganalysis (determining the length of the hidden message). We utilize StegExpose for steganalysis to test the software reliability in hiding messages effectively from steganalysis.

3.6 Performance Speed and Robust Steganography

The straightforwardness of the embedding algorithm has also resulted in the good embedding speed. Most of the files worked with using the software take less than 60 to 90 seconds for embedding. Typically, about 30 seconds is required for decryption. Since the bit per pixel payload is less than 25%, the random number generator does not have to repeatedly struggle to find empty pixels that have not been previously encoded.

4. Software Implementation

The software implementation consists of three executable files: a key generator program, an encryption program, and a decryption program. The encryption program has image analysis functions and windows built-in to aid in cover image categorization.

4.1. Key Generation

A screenshot of the key generation program shown below. The key generation program constructs image keys of random colored pixels according to the user preference for size and file naming. Up to five previously discussed random number generators can be chosen from to generate the random colored pixels.

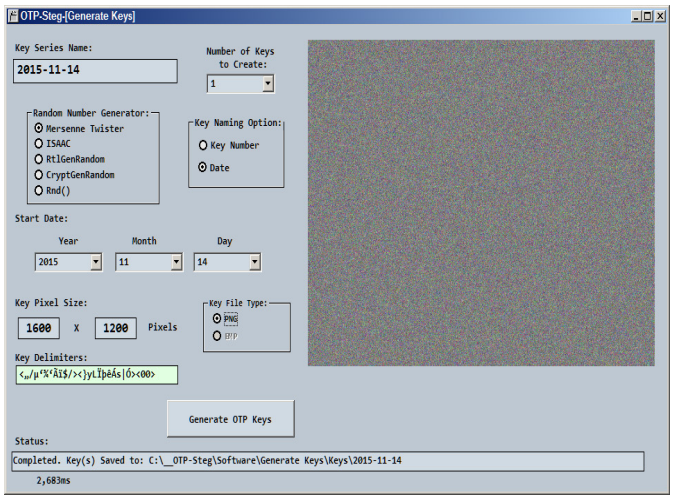


Figure 4. Key Generator executable program

4.2. Encryption and Embedding

The encryption program has by wide margin the most features and functionality built-in. Also included are functions for deleting and forensically wiping the key file used for encryption, as well as the original cover image. By comparing the encrypted payload file to an original cover image, steganography could easily be detected as the difference between the two images. It is an extremely important security measure to eliminate the original cover image and key as soon as possible after encryption takes place.

4.3. Message Hash Value

SHA-256 values of the message are calculated in both the encryption and decryption steps. In the encryption step, the hashed value is incorporated into the end of the message string. Upon decryption, the transmitted hash value is compared to the hashed value of the decrypted message, and displayed in the decryption program graphical user interface. If the values match the user is informed that the message has not been altered in any way since it was encrypted by the sender. This is also a double check that successful decryption has taken place and the message is authentic.

4.4. Text Compression:

Compression prior to embedding the message generally reduces message size by 50 to 80%. The zLib compression library DLL is utilized and called as a function within both the encrypt and decrypt programs. The result makes encryption and decryption quicker and also has the benefit of reducing the bit per pixel payload size in the cover image, increasing security against detection.

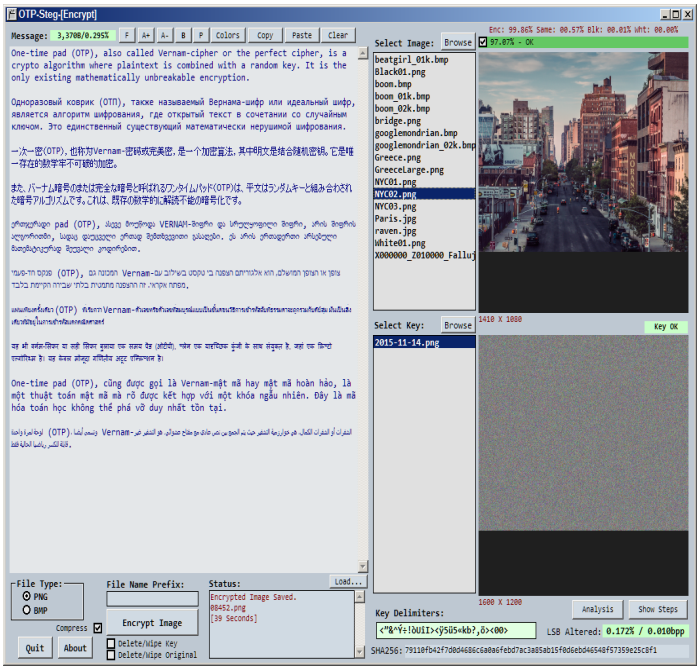


Figure 5. Encrypt/Embed executable program

4.5. Cover and key File Deletion and forensic Data Wiping

File wiping utilities are used to delete individual files from an operating system mounted drive. The advantage of file wiping utilities is that they can accomplish their task in a relatively short amount of time as opposed to disk cleaning utilities which take much longer and must be run separately.

4.6. Built-In Cover and Stego Image Analysis Tools

Several image statistical analysis features are built into the encryption program. Peak signal-to-noise ratio (PSNR), RS, Chi-Squared, LSB visual analysis, color changes, and color variations. The distortion in the stego-image can be measured by parameters such as mean square error (MSE) and peak signal-to-noise ratio (PSNR) (see Equation 1 and 2 below), and correlation. The lesser distortion means, the lesser MSE, but higher PSNR. If p is a M 譌 grayscale image and q is its stego-image, then the MSE and PSNR values are computed using (1) and (2). For color images a pixel comprises 3 or 4 bytes. Each byte can be treated as a pixel and the same equations can be used to calculate the MSE and PSNR.

$$MSE = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} (p_{ij} - q_{ij})^2 \quad (1)$$

$$PSNR = 10 \times \log_{10} \frac{c_{max}^2}{MSE} \quad (2)$$

The software image analysis window in the encryption program is shown below. Using this window, several operations can be performed to estimate effectiveness of message embedding. Least significant bit (LSB) color values can be investigated visually. Shown on the right of the analysis window are the least significant bit values of the photo on the left. If the least significant bit for red, green, and/or blue is set, this color is added at full intensity to the respective pixel in the image on the right. In Figure 7, individual least significant bit color values can be investigated as well in the red, green, and blue channels. In this image, it is obvious there is a problem with the blue channel — the sky has full intensity for all values. Encoding message data here would be risky, as the pixel variation is nonexistent. Steganography would be very easily detected by any encoding in this area. As a result, the software spreads out the message embedding adaptively, and ignores the blue channel in the area of the sky. Since the red channel has the most random variation throughout the image, it carries the largest brunt of the payload, leveraging its random character throughout the image.

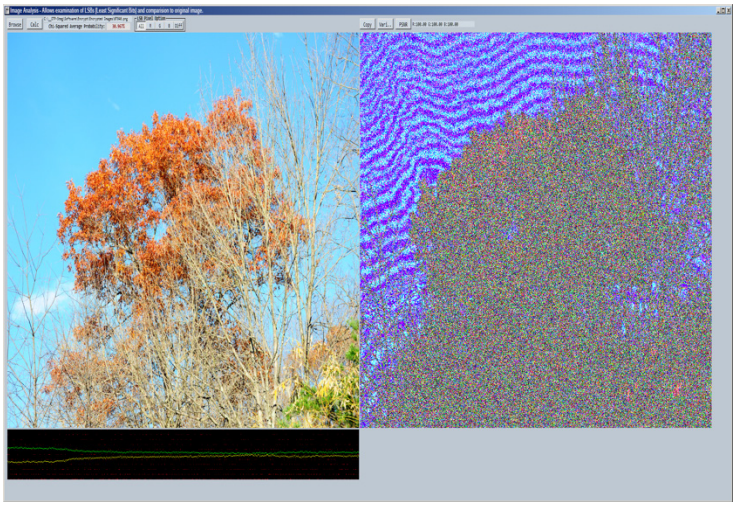


Figure 6. Image Analysis window, cover image on left, LSB analysis on the right

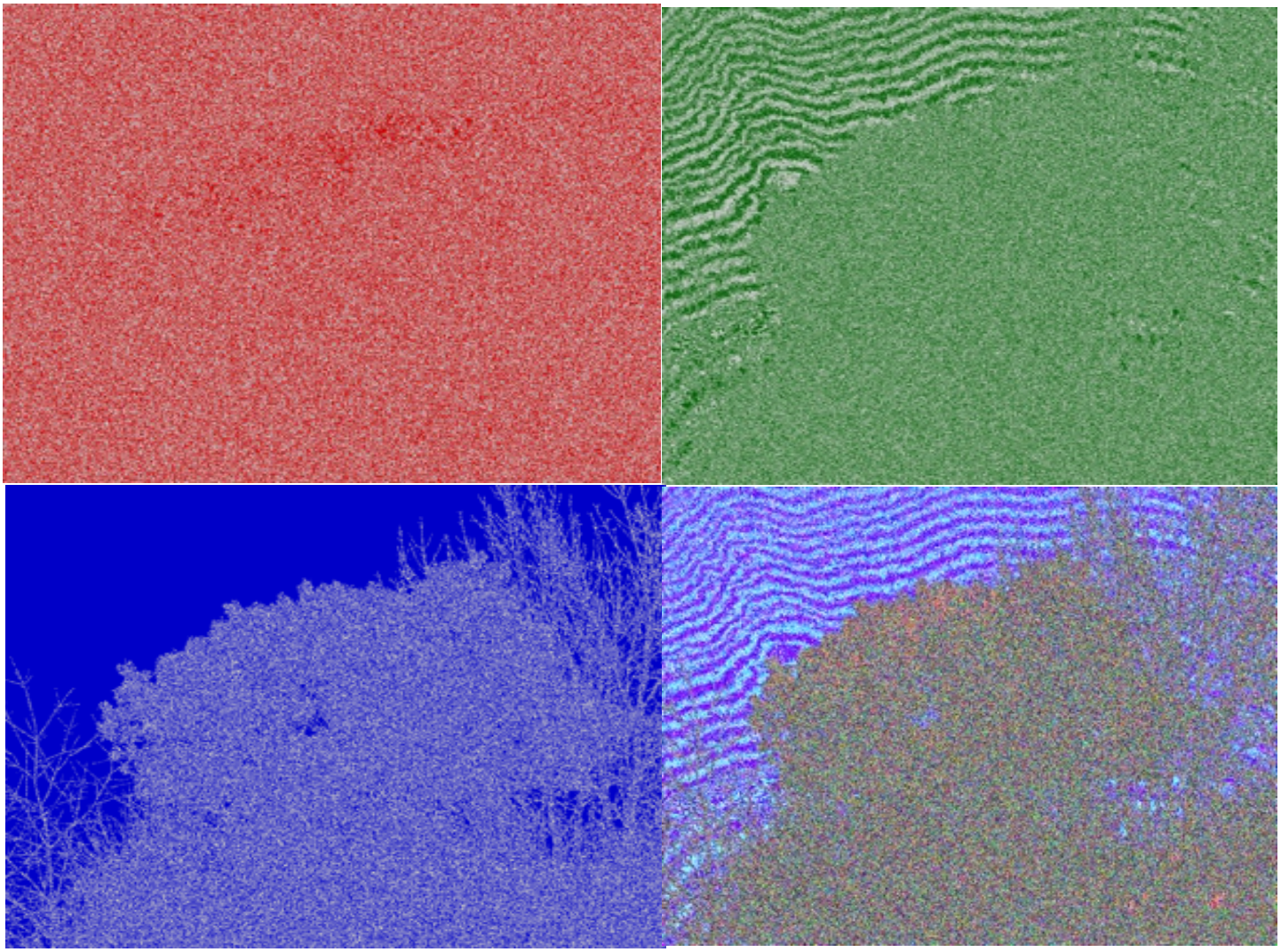


Figure 7. Analysis LSB color analysis graphics (Red, Green, Blue, All Colors).

Shown below are the variations in the red channel, the blue channel in Figure 9 shows the lack of variation in the sky area.

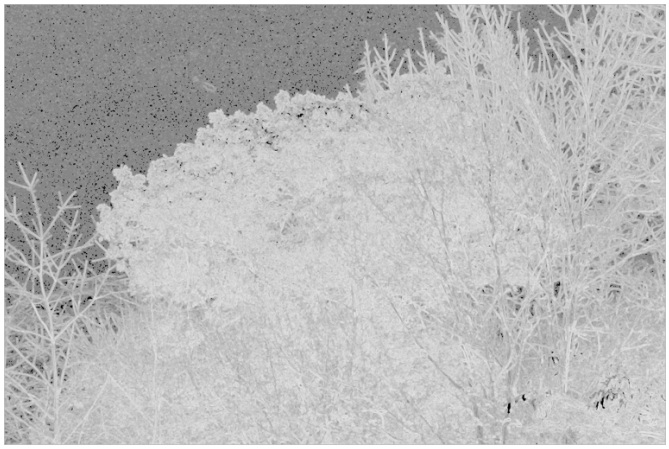
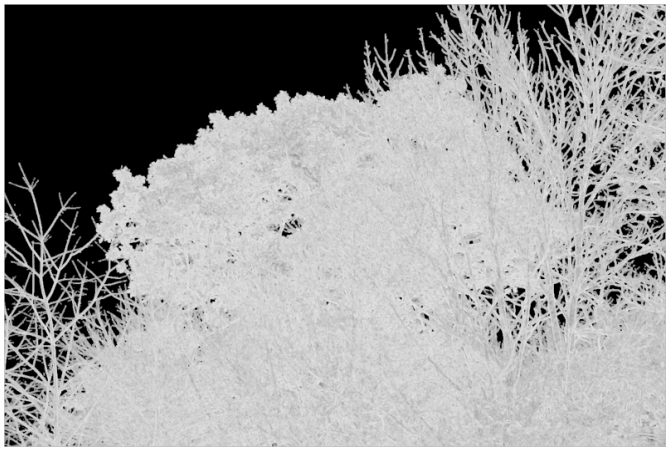


Figure 8. Red channel variation score (normalized to 0-255).



In Figure 10 below, the pixel least significant bit encodings are shown. Notice that in the area of the sky, only red pixels are encoded in the least significant bit. Other areas of the image vary between green and blue embedding depending on which color has the most variation in that pixel general area. Figure 11 shows a blowup of the pixel least significant coding in the area of the transition between the trees and the sky. Notice that the pixel encodings shift from primarily blue-green to the red color at this transition.

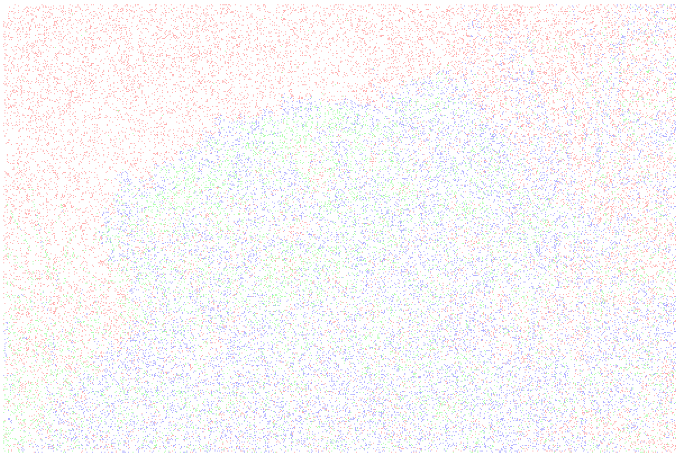
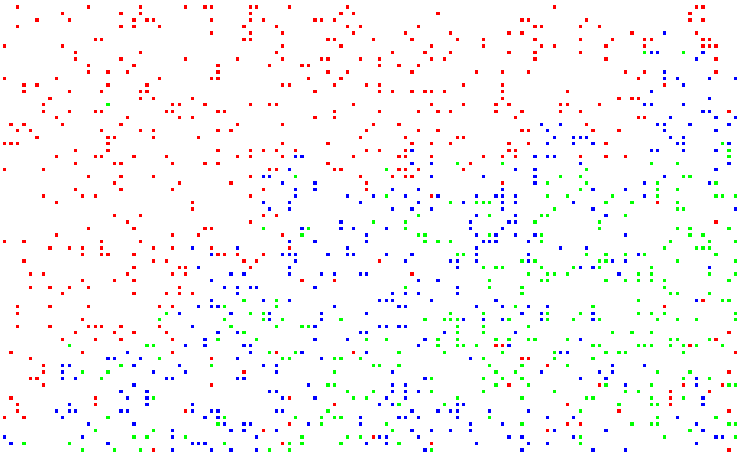


Figure 10. Pixel LSB modification encodings (Red, Green, or Blue).



4.7. *Decryption*

The decryption process largely reverses the encryption process using the decryption program. A SHA-256 hash value is computed from the decrypted message and compared to the hashed value contained within the payload file. If the two values matches, the hashed value is presented with a green colored background. If not, the background is reddish. A green value indicates to the receiver that the message has not been altered in any way since it was written.

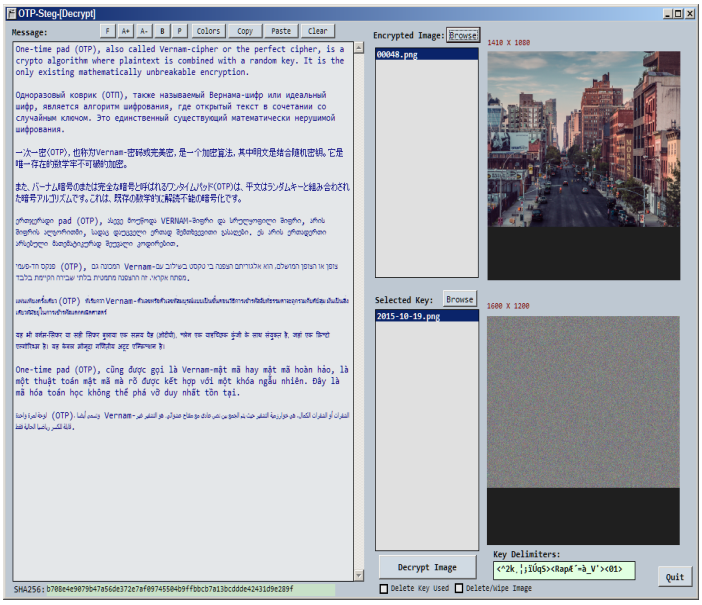


Figure 12. Decrypt/Extract executable program

5.1. *Photo Selection*

There are several general guidelines for photo image selection to increase security. Original photos taken with the user's own camera should be selected as cover images. This is to ensure that the duplicate of the original photo does not exist somewhere on the Internet for comparison. The photo should never have previously been encoded to JPEG to ensure full CMOS pixel sensor color variations throughout the image. As mentioned previously, once

these criteria are satisfied, the user can evaluate an encodability score that is calculated by the encryption program that ranges from 0 to 100%.

5.2. *Encodability Score*

The user should choose in general images that score above 90% for encodability to enhance steganalysis security. The following is the weighting breakdown for the encodability score:

25% Overall PSNR (dB) variation score (0-100%) (more color variation = higher score)

25% Same colors next to each other (0-5%) (less same colors = higher score)

25% Black pixels (0-5%) (less black = higher score)

25% White pixels (0-5%) (less white = higher score)

100% - (100-90% = OK, 90-80% = Marginal, <80% = Unacceptable).

5.3. *Recommended Steganographic Practices*

Table 2. Recommended Steganographic Practices.

No.	Practice	Description
1	Software Operation	Steganography software should be operated on a computer that is not connected to any network or the Internet. Files should be transferred using write only media such as DVD or CD, or less securely by USB drive.
2	Original Photos	Only original photos taken by the users high-quality camera should be considered as cover images. This is to ensure that the image does not have duplicates available on the Internet. Use RAW (original camera file format) images where possible. The software directly accepts all RAW image file types including Nikon, Canon, Sony, etc.
3	Software USB Loaded	The software should be run off of a USB drive plugged into the isolated computer. Further, USB drive containing software, keys, and cover images, should be located separately from the isolated computer in a safe and secure location.
4	Isolated Computers	The isolated computer used to run the software should be well secured and not networked in any way. The operating system should be directly installed from DVD, and antivirus and checks for malware should be regularly run to ensure there is no keystroke loggers, rootkits, or other security compromises installed.
5	"To" and "From" Keys	Both sender and receiver should have their own set of unique keys. Sender A to B, and B to A, each use their own one-way key series. This is to prevent key reuse. Each key must be used only one time, and one time only. Security using OTP depends on this precept.
6	Exchanging Keys	Key exchange should take place upon *physical meeting* using write only media such as DVD or CD. Key exchange must not take place over a network. Keys should be securely generated on isolated computers. Keys must be stored on removable USB drives separate from the isolated computer.
7	Deleting Files	All files including cover image files and key files should be forensically deleted and wiped once used. Forensic wiping utilities in the encryption and decryption programs can be used for this purpose. Wiping consists of randomly overwriting the previous file seven times with random data.
8	Sending Encrypted Files	Encrypted files should be sent as anonymously as possible. Direct email exchange should be avoided. A preferable alternative is to upload files periodically to gallery websites which have potentially thousands of viewers and downloaders daily. Identifying the specific receiver will be difficult in this situation. Each sender should upload to a different anonymous gallery.
9	Monitoring Windows Vulnerabilities	It should be known that just the act of plugging in a USB drive into a Windows computer creates a digital trail throughout the system registry. Installing software using a setup program also creates numerous records within the operating system registry. As a result, the software should be run off of a USB drive without running a separate install/setup utility. Windows must be isolated off of any network to ensure malware is not installed.

10	Malware	Malware can cause a compromise in the steganography system at any time. A keystroke logger that is uploading typed messages is an instant fail. Users must be extremely cautious and knowledgeable about potential malware threats before using the software. In particular, any networked computer presents a point of vulnerability — the software and keys must never be used here. Only transfer of files previously encrypted on an isolated computer can be conducted over a network.
11	Usage Limitations	The biggest limitation is the human factor. Operational security must be observed that all times in addition to technical security. This means aggressive securing of the USB drive use for the software and keys, as well as limited knowledge by parties involved. People should be informed on a need to know basis only.
13	Encrypted Keys	For further security, keys can be encrypted for storage. As a result they will not be able to be used unless the user has knowledge of the encryption key.
14	Wipe Original Photos	Original photos must be deleted and erased from the camera, storage medium, and USB drive as soon as possible after they are used.
15	Wipe Used Keys	Keys must be deleted and erased as soon as they are used.
16	Internet Computer "Clean"	The computer connected to the Internet must be clean of viruses and malware or keystroke loggers. Special care must be taken in this area.
17	Camera Secured	The photo CMOS sensor output profile can be mapped to a particular individual camera. Photo sent on the Internet can be matched up to the users camera. As a result an effort should be made to keep the camera secure.
18	File Upload Galleries	File upload galleries should be selected for anonymity and high traffic volume.
19	Carefully Selected Cover Images	Cover images should be conforming to high encodability statistics and originality. Also they should be of subject matter that will not raise any suspicions.
20	Image File Format	Various image file formats can be chosen, leveraging the fact that steganalysis software will not run on many different types of image file types. Take advantage of other lossless file formats besides PNG and BMP such as TIF, J2K, EXR, WEBP, and JXR.
21	Must-Dos	1: Keep software and keys in secure locations on USB drives. 2: Use software on isolated computer not connected to Internet. 3: Use keys and photos only once and be sure to erase files as soon as possible, especially original cover images and keys.

5.4. Steganographic Communication Security State Level Estimation

We envision certain levels of steganographic communication security levels that correspondents can use for planning, analysis, and security estimations. Thresholds can be established for protective measures using these security level guidelines.

Table 3. Notional Steganographic Security Levels.

Security Level	**Name**	**Description**	**Impact**
10	Secure	Communication commencing securely. Operational security and human threat and insider threat must be strongly monitored and evaluated here.	None-success
9	Communication Suspected	Authorities suspect communication without knowledge of sender and receiver.	Low
8	Steg Statistically Detected	Positive steganography screening results indicating further investigation.	Moderate

7	Internet Computer Searched	If proper security measures recommended previously are followed, nothing should be derived. Duress codeword should be immediately used and communication ceased.	Moderate
6	Transmitted Files Discovered	If proper procedures are used, locating these files should not present much evidence.	Moderate
5	Software Computer Discovered	Traces of software use should be detectable in Windows registry.	High
4	Steg Known	Investigators conclude illicit communication has taken place, without acquiring USB drive(s).	High
	USB Drive discovered	User should make efforts to inform receiver communication is compromised.	Severe
2	Software Discovered/ Acquired	Knowledge of message text should be assumed at this point.	Severe
1	Key(s) acquired	Complete security compromise.	Severe
0	Suspect Detained	All communicating parties should make efforts to destroy any remaining evidence.	Failure

Communicators should have a procedure in place to indicate ceasing of messages and also to destroy related software and keys systematically.

Steganographers should consider incorporating a duress codeword into their communication security protocol. The duress code word should be a predetermined word or phrase that indicates to the receiving party that communication security has been compromised. For example, capture by authorities may have created a situation where one party is succumbing to efforts to be "turned". The duress code word indicates such a situation and should be carefully chosen to arouse no suspicion should authorities have knowledge of its inclusion in a "trap" message.

5.5. Software Availability

The software is available as a free educational and research download to be used for digital forensics education and related projects. Please feel free to use the software for your own educational and research purposes. The software can be acquired here: http://199.175.52.196/OTP-Steg/.

6. Steganalysis Results

StegExpose is a Java based steganalysis tool heavily geared towards bulk analysis of lossless images. It is a steganalysis tool specialized in detecting LSB (least significant bit) steganography in lossless images such as PNG and BMP. It has a command line interface and is designed to analyze images in bulk while providing reporting capabilities and customization which is comprehensible for non-forensic experts. The *StegExpose* rating algorithm is derived from an intelligent and thoroughly tested combination of pre-existing pixel based steganalysis methods Two new fusion detectors, standard and fast fusion were derived from four well known steganalysis methods and successfully implemented in the tool. Standard fusion is more accurate than any of the component detectors it is derived from.

The following LSB steganalysis methods have been incorporated in *StegExpose*. RS analysis (Fridrich, Goljan, and Du 2001) detects randomly scattered LSB embedding in grayscale and color images by inspecting the differences in the number of regular and singular groups for the LSB and "shifted" LSB plane. Sample pair analysis (Dumitrescu, Wu, and Wang 2003) is based on a finite state machine whose states are selected multisets of sample pairs called trace multisets (Dumitrescu, Wu, and Wang 2003). The chi-square attack (Westfeld and Piltzmann 2000) is a statistical analysis of pairs of values (PoV's) exchanged during LSB embedding. PoV's are groups of binary values within a object's LSB's. Primary sets (Dumitrescu, Wu, and Memon 2002) is based on a statistical identity related to certain sets of pixels in an image. The difference histogram analysis (Zhang and Ping 2003) is a statistical attack on an image's histogram, measuring the correlation between the least significant and all other bit planes. Two new fusion detectors, standard and fast fusion, were derived from four well known steganalysis methods and successfully implemented in the tool. The standard fusion test is more accurate than any of the component detectors it is derived from.

StegExpose (the free open source download), was run on a batch of 27 image files that were encoded using the *OTP-Steg* software. Test specifications and results are shown below. None of the embedded files were detectable above the preset default threshold. Standard fusion was the test run which consists of all of the specific steganalysis tests mentioned above.

Table 4. StegExpose Steganalysis Test Specifications.

Test Spec	**Description**
Embedded Text File:	U.S. Constitution; 52,782 Bytes Unicode (422,256 bits)
Images:	27 Various landscape PNG photos, 1200?97 pixels (956,400 pixels) Nikon D90.

Uncompressed:	Approximate Bits per Pixel (bpp) 0.442 bpp
Compressed (zLib):	Approximate Bits per Pixel (bpp) 0.086 bpp
Alterations:	1.445% LSBs altered, 4.335% of pixels
File Archive:	http://199.175.52.196/OTP-Steg/USConstitution/

Table 5. StegExpose Steganalysis Test Results using "Standard Fusion" test.

File name	Above Stego Threshold?	Primary Sets	Chi Square	Sample Pairs	RS analysis	Fusion (mean)
00247.png	FALSE	0.023408176	0.00353364	null	0.02018579	0.01570926
02155.png	FALSE	0.068625394	0.01936033	null	0.04494632	0.04431068
02664.png	FALSE	NaN	5.03E-13	null	0.08658666	0.04329333
03090.png	FALSE	NaN	0.0037011	null	0.23737088	0.12053603
03164.png	FALSE	0.136200359	0	null	0.02282364	0.05300800
03504.png	FALSE	NaN	0.00363950	null	0.19724031	0.10043991
03509.png	FALSE	0.120022314	0.00140079	null	0.03595793	0.05246034
04031.png	FALSE	0.004125309	3.57E-04	null	0.04380402	0.01609549
04095.png	FALSE	NaN	0.00743453	null	0.09919615	0.05331534
04164.png	FALSE	NaN	0.01840689	null	0.07674373	0.04757531
04378.png	FALSE	NaN	4.83E-04	null	0.17958705	0.09003513
04479.png	FALSE	0.047114348	0.00183215	null	0.06152043	0.03682231
04637.png	FALSE	NaN	3.57E-04	null	0.09375770	0.04705742
05169.png	FALSE	0.030743209	3.57E-04	null	0.03714153	0.02274730
05255.png	FALSE	NaN	3.57E-04	null	0.11245120	0.05640417
05262.png	FALSE	0.018022058	0.00206287	null	0.01099853	0.01036115
05777.png	FALSE	0.017279631	6.59E-13	null	0.0070650	0.00811488
06202.png	FALSE	NaN	4.25E-04	null	0.09364117	0.04703301
06672.png	FALSE	0.06420808	0	null	0.06458346	0.04293051
07134.png	FALSE	0.03542274	3.57E-04	null	0.017337435	0.01770577
07140.png	FALSE	NaN	0.00142365	null	0.165817881	0.08362076
07946.png	FALSE	NaN	1.02E-11	null	0.072127587	0.03606379
08145.png	FALSE	0.033316358	2.77E-04	null	0.023061286	0.01888482
09061.png	FALSE	0.014700003	0.00485040	null	0.025382546	0.01497765
09252.png	FALSE	0.074362745	7.14E-04	null	0.01319539	0.02942414
09431.png	FALSE	0.040878552	0.00328135	null	0.031193448	0.02511778
09988.png	FALSE	0.062680713	3.54E-04	null	0.054694774	0.03924330

Test Results Summary: Zero (0%) steganalysis detections using the "Standard Fusion" detection algorithm in *StegExpose* software. *StegExpose* can be downloaded here: https://github.com/b3dk7/StegExpose.

7. Applications Regarding Crypto-Currency Cold-Storage and Transfer

Recently the global phenomenon of crypto-currencies has made headlines and changed the face of financial technology worldwide. The seminal block-chain algorithm has been applied to at least 1200 new crypto-currencies in addition to original, the Bitcoin. However, the future of this revolution is uncertain as security vulnerabilities and regulatory restrictions may make the future of ownership and transfer questionable.

To promote safe usage and encourage investment and speculation for the future, steganography applications such as the one discussed in this paper can be taken advantage of with robustly secure results. For example, Bitcoin public address and private key values can be securely encrypted and hidden in images uploaded to the cloud, transferred through e-mail, or stored locally. This allows applications in so-called cold storage scenarios as well as covert transfer of crypto-currency amounts when necessary.

Transfers using the standard Bitcoin network protocols are easily detectable, and blocked by firewalls and deep TCP/IP packet inspection. As a result, any activity in the clear over the network or protocol whatsoever can be used to flag the IP address and responsible parties--severely compromising one of the precepts of the crypto-currencies which is both anonymity and privacy. VPN traffic has been and is virtually entirely shut down in several entire nations, making this obvious alternative for network usage infeasible.

In the case of Bitcoin, TCP/IP ports 8332 and 8333 are two default ports used for Bitcoin peer-to-peer transfers and can be easily detected, inspected, and blocked, and the information derived used for further investigation into the origin and nature of the network traffic. As a result, it is highly desirable to have secondary means for engaging in transfers other than the standard protocol in situations where deep packet inspection or firewalls may be interfering with the transaction process. This is the case not only for Bitcoin but all of the other crypto currencies as well, now and most likely in the future.

One-Time-Pad Steganography offers this potential in a highly secure manner. Further, in the instance of cold storage with retrieval backup, several individuals can be entrusted with encrypted images and keys, and some of the information can be doubly or triply encrypted to ensure the agreement of several individuals to the release of crypto-currency funds. Encrypted images can be stored on the cloud in multiple locations, as well as locally on permanent storage media. Keys, however, must be stored with the person of the control authority. Additional customized protocols and policies should be designed in the future to address crypto-currency authentication and management; however, the steganography technology offers an ideal starting point for both hidden and theoretically unbreakable storage.

8. Conclusion

In this paper, we have presented a complete One-Time Pad encryption and steganography system including all software necessary to complete practical communication. We have compiled recommended best practices and identified potential security levels. Finally, we have tested the software using robust state-of-the-art steganalysis techniques and found the low payload threshold maintained in the software produces a high margin of communication security safety. No payload files were detected (0% detections), despite each file containing the entire content of the U.S. Constitution as embedded text.

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this paper.

References

[1] R. J. Anderson, and F. A. P. Petitcolas, In the limits of steganography? IEEE Journal of Selected Areas in Communications, vol.16, no.4, pp.474-481, 1998.

[2] M. Bashardoust, G. B. Sulong, and P. Gerami, Enhanced LSB image steganography method by using knight tour algorithm, vignere encryption and LZW compression? International Journal of Computer Science Issues, vol.10, no.2, pp.221-227, 2013.

[3] Bhatacharya, I. Banerjee, and G. Sanyal, A survey of steganography and steganalysis techniques in image, text, audio and video cover carrier? Journal of Global Research in Computer Science, vol.2, no.4, pp.1-16, 2011.

[4] K. Chan, and L. M. Chang, Hiding data in images by simple LSB substitution? Pattern Recognition, vol.37, pp.469-474, 2004.

[5] Cheddad, J. Condell, K. Curran, and P.M. Kevitt, Digital image steganography: survey and analysis of current methods? Signal Processing, vol. 90, pp.727-752, 2010.

[6] S. M. Douiri, M. B. O. Medeni, S. Elbernoussi, and E. M. Souidi, A new steganographic method for gray scale image using graph coloring problem? Applied Mathematics & Information Sciences, vol.7, no.2, pp.521-527, 2013.

[7] Gangwar, and V. Srivastava, Improved RGB-LSB steganography using secret key? International Journal of Computer Trends and Technology, vol.4, no.2, pp.85-89, 2013.

[8] R. S. Gutta, Y. D. Chincholkar, and P. U. Lahane, Steganography for two and three LSBs using extended substitution algorithm? ICTAT Journal on Communication Technology, vol.4, no.1, pp.685-690, 2013.

[9] Gutub, M. Ankeer, M. Abu-Ghalioun, A. Shaheen, and A. Alvi, Pixel indicator high capacity technique for RGB image based steganography? in Proceedings of Fifth IEEE International Workshop on Signal Processing and its Applications, 2008, University of Sharjah, U.A.E.

[10] N. Hamid, A. Yahya, R.B. Ahmad, D. Nejim, and L. Kannon, Steganography in image files: a survey? Australian Journal of Basic and Applied Sciences, vol.7, no.1, pp.35-55, 2013.

[11] J. He, S. Tang, and T. Wu, In adaptive steganography based on depth-varying embedding? in Proceedings of 2008 Congress on Image and Signal Processing, 2008, pp.660-663.

[12] M. Hussain, and M. Hussain, A survey of image steganography techniques? International Journal of Advanced Science and Technology, vol. 54, pp.113-123, 2013.

[13] Y. K. Jain, and R. R. Ahirwal, A novel image steganography method with adaptive number of least significant bits modification based on private stego-keys? International Journal of Computer Science and Security, vol.4, no.1, pp.40-49, 2010.

[14] M. Juneja, and P.S. Sandhu, Designing of robust image steganography technique based on LSB insertion and encryption? in Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, 2009, pp.302-305.

[15] Kamaldeep, Image steganography techniques in spatial domain, their parameters and analytical techniques: a review article?, IJAIR, vol.2, no.5, pp.85-92, 2013.

[16] H. B. Kekre, A. A. Athawale, and P. N. Halarnkar, Increased capacity of information hiding in LSB's method for text in image? International Journal of Electrical, Computer and System Engineering, vol.2, no.4, pp.246-249, 2008.

[17] Y. K. Lee, G. Bell, S.Y. Huang, R.Z. Wang, and S.J. Shyu, In advanced least-significant-bit embedding scheme for steganographic encoding? LNCS, vol.5414, 2009, pp.349-360.

[18] Li, J. He, J. Huang, and Y.Q. Shi, A survey on image steganography and steganalysis? Journal of Information Hiding and Multimedia Signal processing, vol.2, no.2, pp.142-172, 2011.

[19] C. Lou, and C. H. Hu, LSB steganographic method based on reversible histogram transformation function for resisting statistical steganalysis? Information Sciences, vol.188, pp.346-358, 2012. Application of a large key cipher in image steganography by exploring the darkest and brightest pixels? International Journal of Computer Science and Communication, vol. 3, no.1, pp.49-53, 2012.

[20] R. S. Marcal, and P.R. Pereira, A steganographic method for digital images robust to RS steganalysis? LNCS, vol.3656, 2005, pp.1192-1199.

[21] Martin, G. Sapiro, and G. Seroussi, Is image steganography natural? IEEE Transactions on Image Processing, vol.14, no.12, pp.2040-2050, 2005.

[22] M. K. Meena, S. Kumar, and N. Gupta, Image steganography tool using adaptive encoding approach to maximize image hiding capacity? International Journal of Soft Computing and Engineering, vol.1, no.2, pp.7-11, 2011.

[23] Mishra, A. Gupta, and D. K. Vishwakarma, Proposal of a new steganography approach? in Proceedings of International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009, pp.175-178.

[24] H. Mathkour, G. M. R. Assassa, A. A. Muharib, and I. Kiady, A novel approach for hiding messages in images? in Proceedings of International Conference on Signal Acquisition and Processing, 2009, pp.89-93.

[25] H. Motameni, M. Norouzi, and A. Hatami, Labeling method in steganography? World Academy of Science, Engineering and Technology, vol. 24, pp.349-354, 2007. vol. 270, part II, 2012, pp.479-488.

[26] M. T. Parvez, and A. A. Gutub, RGB intensity based variable-bits image steganography? in Proceedings of IEEE Asia-pacific Services Computing Conference, 2008, pp.1322-1327. Gandharba Swain et al. / International Journal of Computer Science & Engineering Technology (IJCSET)

[27] P. S. Pharwaha, Secure data communication using moderate bit substitution for data hiding with three-layer security? IE(I) Journal-ET, vol.91, pp.45-50, 2010., International Journal of Security and Its Applications, vol.6, no.2, pp.1-12, 2012.

[28] G. Swain, and S. K. Lenka, RSB array based image steganography technique by exploring the four least significant bits? CCIS

[29] G. Swain, D. R. Kumar, A. Pradhan, and S. K. Lenka, A technique for secure communication using message dependent steganography? International Journal of Computer and Communication Technology, vol.2, no. 2- 4, pp.177-181, 2010.

[30] G. Swain, and S. K. Lenka, Steganography using the twelve-square substitution cipher and an index variable? in Proceedings of ICECT, 2011, vol.3, pp.84-88.

[31] G. Swain, and S. K. Lenka, A robust image steganography technique using dynamic embedding with two least significant bits? Advanced Materials Research, vols. 403-408, pp.835-841, 2012.

[32] G. Swain, and S. K. Lenka, A dynamic approach to image steganography using the three least significant bits and extended hill cipher? Advanced Materials Research, vols. 403-408 pp.842-849, 2012.

[33] G. Swain, and S. K. Lenka, A technique for secret communication by using a new block cipher with dynamic steganography"

[34] G. Swain, and S. K. Lenka, A hybrid approach to steganography- embedding at darkest and brightest pixels?, in Proceedings of International Conference on Communication and Computational Intelligence, 2010, pp.529-534.

[35] M. A. B. Younes, and A. Jantan, A new steganography approach for image encryption exchange by using least significant bit insertion, International Journal of Computer Science and Network Security, vol.8, no.6, pp.247-254, 2008.

[36] H. J. Zhang, and H. J. Tang, "A novel image steganography algorithm against statistical analysis", in Proceedings of Sixth International Conference on Machine Learning and Cybernetics, 2007, pp.3884-3888.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 276-283 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

The Impacts of Distribution Generators' Size and Location on Power Efficiency and Voltage Profile in Radial LV Networks

Maher M.A Al-Maghalseh*,1,2

[1]*Electrical Engineering Department, Palestine Polytechnic University, Palestine*

[2] *Renewable Energy and Enviermental Research Unit (REERU), Palestine Polytechnic University, Palestine*

ARTICLE INFO

Article history:
Received: 20 June, 2018
Accepted: 31 July, 2018
Online: 10 August, 2018

Keywords:
Distribution generations (Distribution generators)
Off Grid systems
Grid connected systems
Voltage profile
Power losses

ABSTRACT

This paper presents analytical methods for optimum allocation of distribution generators in a radial distribution network. The aim is to increase power efficiency and reduce loss of power. The ideal size of each distribution generator unit at the identified locations has been carefully selected by making adjustments to loss saving equations and the voltage on a radial feeder. The experimental system is located in the governorate of Bethlehem specifically within the district served by the Jerusalem Distribution Electricity Company (JDECO). The NEPLAN simulator has been utilized in the verification process of the aforementioned experimental system. It entails of several nodes and 400 V-buses that feed up to five household loads. The size and location of each has been carefully chosen. Consequently, significant loss saving, and voltage profile improvement have been reported. Hence, the use of Renewable Distribution Generators (DGs) systems is highly desirable for various reasons including increased system reliability, decreased loss of power, and enhanced voltage profile. Notably, cutting back on the use of conventional fuel will in turn reduce the negative environmental impacts.

1. Introduction

Electric supply companies incur losses of supply on its way to final consumers mainly manifested in the quantities of power infused into distribution grids and the transmission which users are not charged for. Mainly there are two main types of losses: technical and non-technical. This paper will address the technical losses which are losses that occur naturally mainly due to electricity dissipation in power system machineries such as transmissions and distribution channels, transformers, and measurement systems. The losses in electricity are quite higher at the distribution lines compared to that at the transmission systems. This is usually due to the R/X ratio which is higher at the distribution lines. Consequently, reducing the technical losses will increase technical and cost efficiency while relieving the community from the economic and environmental costs of power generation.

Several methods have been implemented in order to save power at distribution systems. Some of which were network reconfiguration and capacitor placement. Notably, the staggering growth of distribution generation may contribute to both the increase in power losses as well as the reduction in power quality at the distribution system. On the other hand, distribution generators, if managed efficiently, can enhance the system's performance.[1]. This paper is an extension of work previously presented in 2017 [2]. This paper presents a thorough technical and economic analysis of standalone and grid connected photovoltaic (PV) systems. A considerable amount of literature has already been published [3] on studies that showcased the latest technologies in distribution reconfiguration. Optimization of technical losses in electricity has sure been deemed an engineering issue that includes designing the system in addition to planning and modeling.

Among several techniques implemented were selecting the distribution generators' optimum locations and size. To that end, Adefarati and Bansal have proposed the Markove model to evaluate the reliability of the distribution system with a wind turbine generator (WTG), an electric storage system (ESS), and a photovoltaic (PV) [4]. Several case studies have been conducted to

*Corresponding Author: Maher M.A Al-Maghalseh, maher4xp@yahoo.com

assess reliability,it was found that using distribution generators does in fact improve the reliability of the distribution system.

Moreover, researchers have found that embedded generation on distribution systems can increase the fault level in the system. Hence, Orozco-Henao et al. [5] have presented a fault location (FL) analytical methodology for active distribution networks. The suggested technique has been used to assess the ideal fault level and location. The ATP/EMTP software has been utilized in modeling the test feeder. The results have emphasized the accuracy of this methodology and its potential for success in real-life applications. Additionally, Kaveh et al. [6] have performed a new analytical method to assess the influence of distribution generators.. It was found that the proposed method can be used to specify the ideal location to set the distribution generator to thus reduce phase imbalance, increase power efficiency, and enhance the voltage profile. Another study by Soroudi et al. [7] has presented analytical methods to assess the effects of wind power on the distribution networks. The proposed methods were applied on a distribution network, namely a 69 bus. Interestingly, the numerical method can be used to specify the optimum location, the capacity of renewable energy sources, and to minimize the power losses.

Pukhrem et al. [8] have studied the effects of Rooftop Solar PV Installation in a low-voltage distribution network. To that end, reactive power control and active power curtailment control methods have been applied to reduce voltage fluctuation into the distribution system. The result of merging both approaches has indicated a positive impact on managing and regulating the voltage profile of the system. A numerical study developed by Matlab Simulink [9] to study the influence of the distribution network on the annual energy yield of residential PV systems with variables of increased PV penetration. The study has been conducted upon actual residential PV systems within the UK distribution network. The results have revealed that the level of PV generation would reduce should network voltage rise beyond grid limits. Additionally, the percentage of the annual energy yield, and the economic performance of the systems have been analyzed.

In [10], the author presented the Monte Carlo-based method to analyse the effects of Rooftop photovoltaic (PV) on voltage quality. While in [11] conducted a numerical study of hybrid photovoltaic (PV) and wind turbine (WT) rooftop generation systems (hybrid systems).it has been found that enhanced penetration level of the hybrid system in turn results in maintaining the voltage within range. In [12], the researcher presented an economic analysis of the installation of PV generation on-site (self-consumption) compared to generating solar energy and exporting it to the grid. The study has been conducted on 302 households which have taken part in a UK smart grid project. It was found that 45% of self-consumption can be supplied by the PV generation. This has resulted in decreasing the average annual electricity demand which amounts to 24 percent reduction. In [13], the author provided a review of high penetration of photovoltaic micro-generation on a low voltage distribution network. The study proposed an analytical method to study the effects of high penetration of PV on the radial distribution network. It has been indicated that 'self-consumption with storage' is the most efficient method; to reduce overvoltage, to prevent reverse power flow, and in turn reduce power losses.

A noteworthy analysis and discussion on the subject has been presented by Al-Maghalseh [14-16] on several real-life case studies of embedded distribution generations. Interestingly, the effects of different types of renewable energy distribution generation have been clearly evaluated. Horowitz et al. [17] provided an extensive review on the cost of integrating PV systems into distribution networks. Omar and Mahmoud [18] explored grid connected PV home systems in Palestine. A significant amount of published literature has discussed the effects of high PV penetration on low voltage distribution networks [19]. Furthermore, a detailed review of PV in the distribution systems has been carried out in [20-22].

This paper exhibits an ideal blueprint of standalone and grid connected photovoltaic systems which cover the networks' demand. The blueprint specifies the ideal size and location of the distribution generator at which it will incur slightest technical losses while enhancing the voltage profile.

2. Proposed Methodology

Recently, many studies have been used to optimize the allocation and size of distribution generators. Different analytical approaches have been applied by [1, 23, 24]. Mixed integer non-linear programming method has been used by [25-28]. Atwa and Saadany have used the evolutionary algorithms (EA) technique [28]. Recently, an alternate method has been adopted for the optimization process [19]. This approach has utilized the equivalent current injection technique and neglected the Jacobian matrices. Furthermore, this approach is in line with the analytical method presented in [1]. hence, it is essential to examine the effects of the reactive loss and the real power loss in the system.

In this paper, a simplified analytical method is applied to examine the distribution generation active and reactive branch currents and the related loss savings of a distribution generation located in the radial distribution system. Figure 1 exhibits the N-bus radial distribution system.

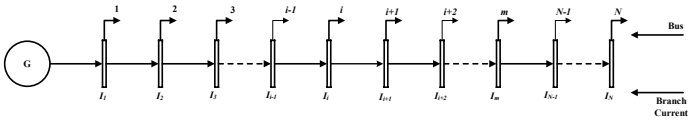


Figure 1. An N-bus radial distribution network.

The total power loss, P_L, of N-bus radial DS having n number of branches, (Figure 1) can be presented as:

$$P_L = \sum_{i=1}^{N-1} I_i^2 . R_i = \sum_{i=1}^{N-1} \left(I_{ai}^2 + I_{ri}^2 \right) . R_i \quad (1)$$

whereas I_i is the complex current, I_{ai} and I_{ri} are the real and imaginary parts of $I_i=(I_{ai} + jI_{ri})$ respectively, while R_i is the resistance of the i_{th} branch. The value of I_i can be obtained from the load flow solution of the given DG.

Figure 2 exhibits the N-bus radial distribution system with a DG. Upon DG placement in the system, the current (I_i^{new}) for the branch i, at m_{th} bus can be presented by the following illustration:

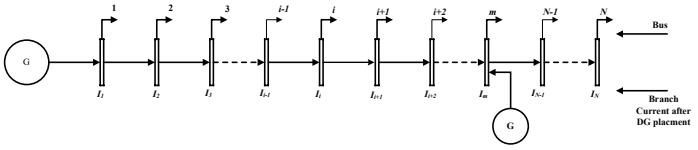


Figure 2. N-bus radial distribution network after distribution generation locating.

$$I_i^{new} = I_i - D_i.I_d g^m = (I_a i - D_i.I_d ga^m) + j(I_r i - D_i.I_d gr^m) \quad (2)$$

whereas,

$$I_i = I_{ai} + jI_{ri} \quad (3)$$

$$I_{dg}^m = I_{dga}^m + jI_{dgr}^m \quad (4)$$

Infusing a current I_{dg}^m influences the currents along the network branches in the area between the substation and the m_{th} bus, while the currents within the rest of the branches remain unaffected. However, I_{dga}^m and I_{dgr}^m are the real and imaginary components of I_{dg}^m and D_i can be illustrated as follows:

$$D_i = \begin{cases} 1; \text{ if branch } i \text{ is between bus 1 and } m \\ 0; \text{ otherwise} \end{cases} \quad (5)$$

To that end, the formula for the total compensated losses P_L^{new} after DG placement may be conveyed as:

$$P_L^{new} = \sum_{i=1}^{N-1} \left(I_i^{new}\right)^2 .R_i = \sum_{i=1}^{N-1} \left[\left(I_{ai} - D_i.I_{dga}^m\right)^2 + \left(I_{ri} - D_i.I_{dgr}^m\right)^2 \right].R_i \quad (6)$$

The related power losses saving, S_L, can be exhibited using (1-6) as:

$$S_L = P_L - P_L^{new} = \sum_{i=1}^{N-1} [\left(I_{ai}^2 + I_{ri}^2\right) - \{\left(I_{ai} - D_i.I_{dga}^m\right)^2 + \left(I_{ri} - D_i.I_{dgr}^m\right)^2 \}].R_i \quad (7)$$

By simplifying the above equation, we arrive at:

$$S_L = 2\sum_{i=1}^{N-1} \left[I_{ai}.I_{dga}^m - I_{ri}.I_{dgr}^m\right].D_i. - \sum_{i=1}^{N-1} \left[\left(I_{dga}^m\right)^2 + \left(I_{dgr}^m\right)^2\right].D_i^2.R_i \quad (8)$$

The highest loss saving is due to the DG currents I_{dga}^m and I_{dgr}^m and can be arrived at using (8) given that the following prerequisites are fulfilled:

$$\frac{\partial S_L}{\partial I_{dgn}^m} = 0 \quad (9)$$

As such, based on equation (9) we can derive :

$$\frac{\partial S_L}{\partial I_{dgn}^m} = 2.\sum_{i=1}^{N-1} I_{ri}.D_i.R_i - 2\sum_{i=1}^{N-1} I_{dga}^m.D_i^2.R_i = 0 \quad (10)$$

$$\frac{\partial S_L}{\partial I_{dgr}^m} = 2.\sum_{i=1}^{N-1} I_{ri}.D_i.R_i - 2\sum_{i=1}^{N-1} I_{dgr}^m.D_i^2.R_i = 0 \quad (11)$$

Upon simplifying equations (10) and (11), the active distribution generation current I_{dga}^m and reactive current I_{dgr}^m representing the highest loss saving formula can be respectively depicted as follows:

$$I_{dgn}^m = \frac{\sum_{i=1}^{N-1} D_i.I_{ai}.R_i}{\sum_{i=1}^{N-1} D_i^2.R_i} \quad (12)$$

$$I_{dgr}^m = \frac{\sum_{i=1}^{N-1} D_i.I_{ri}.R_i}{\sum_{i=1}^{N-1} D_i^2.R_i} \quad (13)$$

The subsequent DG size is arrived at from equations (11-13):

$$S_{dg}^m = V_m.\left(I_{dg}^m\right)^* = V_m.\left(I_{dgn}^m - I_{dgr}^m\right) \quad (14)$$

where, S_{dg}^m is the capacity of m_{th} DG in complex form, and V_m is the phasor voltage at the bus.

3. The Test System

Using NEPLAN software the presented analytical methods have been utilized on a real-life test radial distribution system, for the purpose of optimal DG allocation and sizing. The selected sample involves a system utilized by the Jerusalem District Electricity Co. (JDECO) in the vicinity of the city of Bethlehem. A balanced constant load radial distribution system particularly of a single source (see Figure3). Voltage limitation has been set to a range between -5% and +5%. The DG has been placed south facing the PV systems at a tilted angle of 15° degrees. The type of systems utilized include standalone and grid connected without the effect of shading. Various setups for positioning the DGs have been taken into account. To decrease loss, several possible solutions have been carefully analyzed. Furthermore, a detailed analysis along with a simulation have been conducted for the radial configuration while changing the variables as deemed necessary. The blueprint of the electric network for the aforementioned site is illustrated in Figure3.

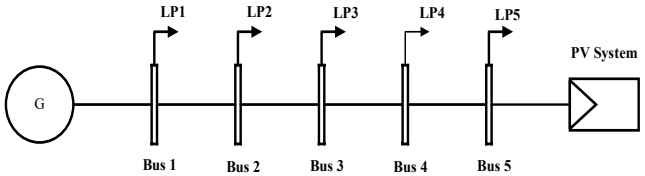


Figure 3. A schematic diagram for the test radial distribution network.

Figure 4. below depicts consumers power consumption in kWh/day. Furthermore, the yearly sun light hours for each month are shown in Figure5.

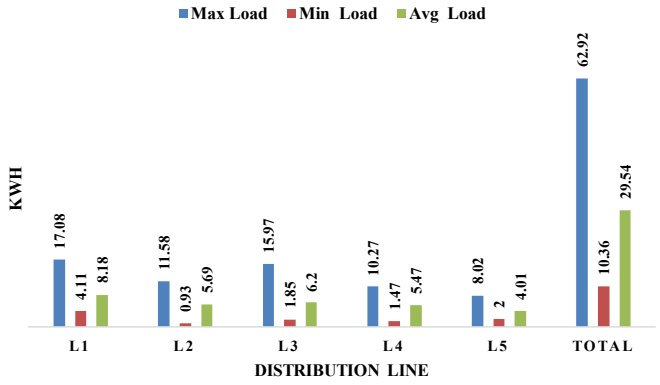


Figure 4. The loads at the distribution line.

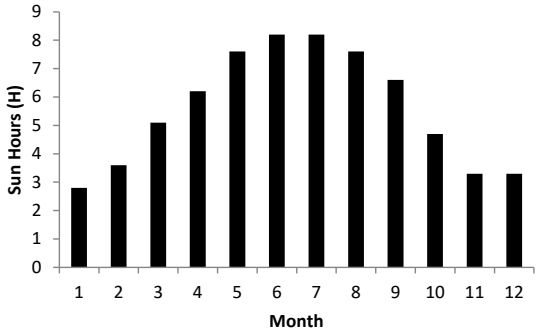


Figure 5. Yearly Sun light hours.

3.1. Photovoltaic system design

The methodology of this study has included researching inputs and the required calculations to determine the potential for either types of solar photovoltaic generation according to the amount of solar radiation.

Grid connected system design:

The electrical characteristics for the panels and the inverter used in the grid connected system are exhibited below in Table 1 and Table 2, respectively.

Table.1. Electrical characteristics for the PV panels under Standard Test Conditions (STC).

Nominal Maximum Power (Pm) in Watts	315
Power tolerance	0 / + 5 W
Open Circuit Voltage (Voc) in Volts	45.25
Short Circuit Current (Isc) in Amps	9.29
Voltage at Maximum Power (Vmp) in Volts	36.75
Current at Maximum Power (Imp) in Amps	8.58
Maximum System Voltage in Volts	1000
Module Efficiency (%)	16.25

Table.2. Technical characteristics of 10 kW inverter

Inverter Type	AS-IC01-xxx-2 (2 MPPT)
Recommended Max. DC voltage (V)	1000
Maximum efficiency	98.30%
MPPT range (V)	200-800
Max. DC power (W)	10000
Max. input current (A)	12.5 x 2
Rated output power (W)	9800
Max. AC output current (A)	14

The system has been designed based on actual figures collected from average power consumption per day. Accordingly, recorded inputs were as follows the consumed energy at an average of 29.54 kWh per day, while sun peak hour approximately 5.58H, whereas the system losses reached approximately 30%, and the future expansion estimated at 20%. Hence, the solar module capacity can be calculated using the equation below as follows:

$$\begin{aligned} Energy\ per\ year &= 29.54\ \text{k}\,WH * 365\ day \\ &= 10,783.65\ \text{kWH / year} \end{aligned} \tag{15}$$

$$\begin{aligned} Solar\ module\ capacity &= \frac{Energy\ per\ year}{365 * \text{SPH} * \left(1 - SYS_{Loss}\right)} \\ &= \frac{10,783.65}{365 * 5.58 * 0.7} = 7.75\ \text{k}\,W \end{aligned} \tag{16}$$

where, SPH stands for sun peak hours

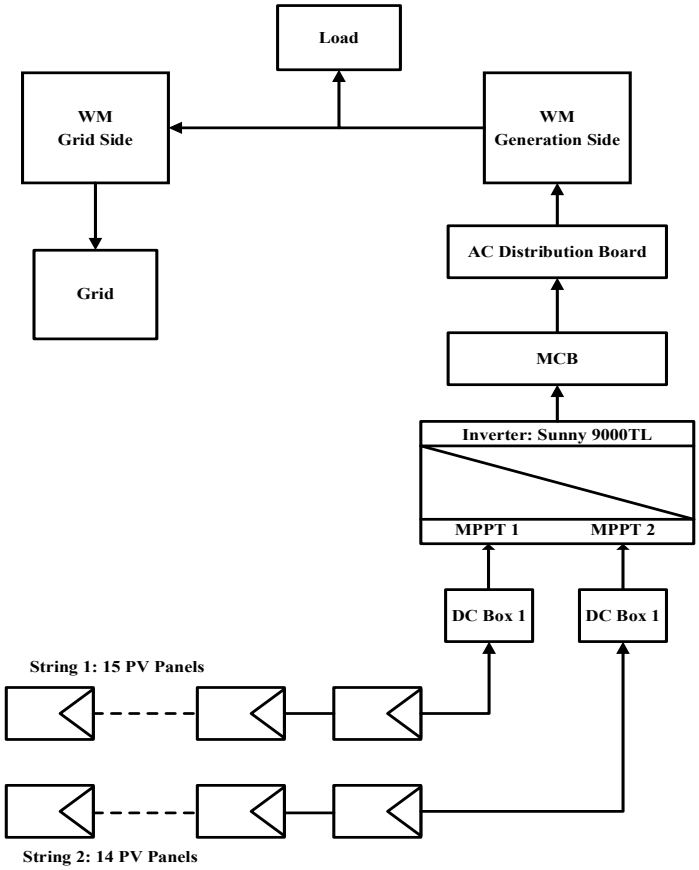


Figure 6. A schematic diagram for the grid connected system.

Adding +20% for future and load growth, the number of modules and energy per year can be calculated using the equation below as follows:

$$\text{Required Capacity} = 7.75 * 1.2 = 9.08\ \text{k}\,W \tag{17}$$

$$\text{Number of module} = \frac{9.08}{315} = 28.57 = 29\ Panels \tag{18}$$

$$\text{Actual system capacity} = 29 * 315 = 9135\,W \tag{19}$$

$$\text{Energy/year} = = \frac{\textit{Actual system capacity} * 365 * \text{SPH}}{\left(1 - SYS_{Loss}\right) * 1000} = \frac{9135 * 365 * 5.58}{0.7 * 1000} = 13,016.93 \text{ kWh} \quad (20)$$

The schematic diagram for the grid connected system design is shown in Figure 6. It exhibits 15 panels connected at input A of the inverter, and 14 panels connected at input B.

Standalone System Design:

The design has been based on the average consumption per day. The energy was recorded at 29.54 kWh per day, the sun peak hour at approximately 2.79H, system losses have reached about 30%, whereas the battery bank can provide up to two autonomy days upon having three sunny days. Panels and an inverter similar to those applied in the grid connected system have been used for the standalone system. Using equation 14, the value of solar module capacity has been approximately 15.11 kW, while accounting for autonomy days.

$$\textit{autonomy factor} = x + \frac{2}{3}x = \frac{5}{3}x = 1.667\,x \quad (21)$$

Consequently, the capacity of the solar system is calculated using equations (15-18) which has been estimated at 25.08 kW. Since the required capacity is approximately 25.08 kW, the estimated number of PV modules needed is 80 Panels. Noting that the system capacity mounted to approximately 25.2 kW, while the general energy generated per year estimated at 17,986 kWh. Hence, to calculate the inverter inputs, we have assumed that: input A= 3 x 16 string panels, and input B = 2 x 16 string panels. In addition, two island inverters have been used. An island inverter to incorporate the storage systems into the standalone systems. Hence, DC voltage for the batteries has been estimated at 48 V. The storage energy for up to two days is calculated using the following formula:

$$Es = \frac{\textit{Generated energy per year}}{365} \times 2 = \frac{17,986.13}{365} x2 = 98.55 \text{ kWH} \quad (22)$$

To assess the needed number of DC batteries, we have assumed using 24V/ 100A batteries. For energy safety:

$$E_{safe} = \frac{E_{rough}}{MDOD} = \frac{98}{0.75} = 130.67\,KWH \quad (23)$$

To estimate the capacity of the battery bank needed the following formula has been used:

$$C = \frac{E_{safe}}{V_b} = \frac{130.67}{24} = 2722\ Amps.h \quad (24)$$

The total number of batteries has been thus obtained using the following formula:

$$N_{batries} = \frac{C}{C_b} = \frac{2722}{100} = 28\ Batteries \quad (25)$$

The number of batteries in each series has been determined by the following formula:

$$N_s = \frac{48}{24} = 2 \quad (26)$$

Finally, the number of parallel paths N_p has been obtained by the following formula:

$$N_p = \frac{N_{batries}}{N_s} = \frac{28}{2} = 14 \quad (27)$$

The number of batteries required is 28 batteries (14 parallel branches and 2 series of batteries). In Taking into account the estimated sun peak hours of 2.79, the energy required for 2 days can be verified as follows:

$$H = \frac{\text{energy required for 2 days}}{\textit{No. of inverters} * \textit{inverter capacity} * \textit{No. of off days}} = \frac{98.55}{2 * 6 * 3} = 2.74 < 2.79 \quad (28)$$

The system has indeed been able to charge batteries to feed the loads for 2 days. Figure7 exhibits the schematic diagram for the standalone system design.

4. Simulation results

In order to select the ideal location to place the DG, the NEPLAN simulator was utilized in testing four case scenarios. Moreover, DG buses have been designed as PQ bus with a power factor of less than one. Listed below are the abovementioned tested case scenarios respectively;

1. The impact of highest load with lowest generation
2. The impact of highest load with maximum generation
3. The impact of lowest load with lowest generation
4. The impact of lowest load with highest generation.

Figure 8 below exhibits a schematic of the designated locations of the DG systems in cases 2 and 4.

Figure9 below exhibits the voltage profiles for cases 1 to 4. According to the voltage profile results, it is evident that mounting the PV system at buses 7 and 5 enhances the voltage levels at all buses thus maintaining a range of ±5%. The voltage profiles for cases 3 and 4.

According to the figure, it is evident that the buses increase in voltage levels is recorded at 1.01 Pu and 1.02 Pu. it is highly unlikely for this case to occur, and even if it were to occur during daylight hours, the voltage range will fall within normal rates. Mounting the DG nearby the loads has resulted in quite enhanced voltage profile This indicates that mounting the DG close to the loads enhances the overall voltage profile curve. To conclude, mounting the DG at bus 4 ensures the lowest energy losses and improves the voltage profile thus keeping it within a normal span.

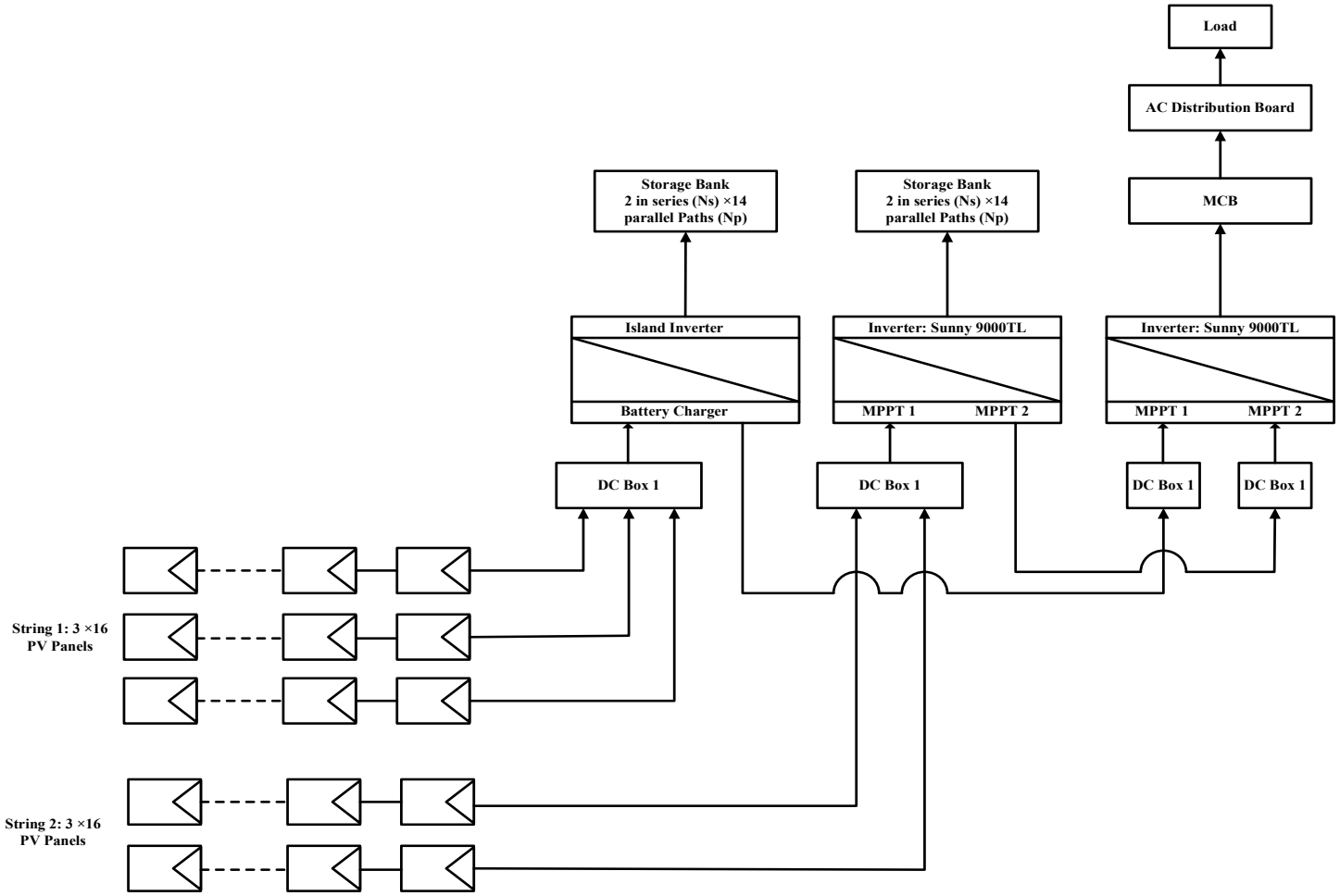


Figure 7. A schematic diagram for grid connected system.

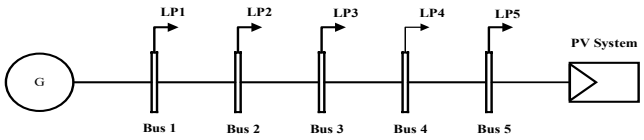


Figure 8. designated locations for the DGs.

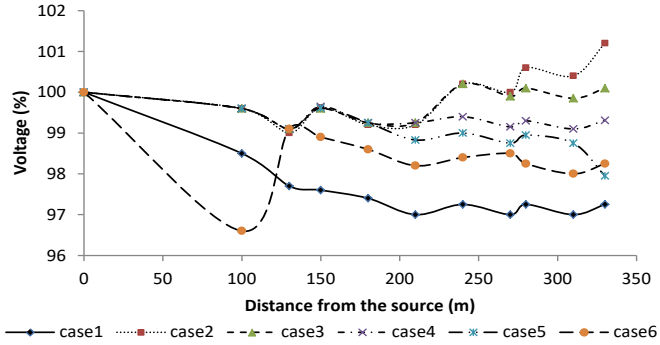


Figure 9. Voltage profiles of the system

In this section, an economic analysis for both standalone and grid connected systems will be featured along with a feasibility evaluation for each option. The evaluation methodology has been based on some financial principles such as: payback period, internal rate of return (IRR), the levelized cost of electricity (LCOE), and the net present value (NPV).

Payback is among widely used methods to estimate the value of a project. Its value is determined by comparing the cumulative cash investments against the cumulative benefits. The payback period formula is calculated by dividing the initial investment over the cash inflow per period:

$$Payback\ Period = \frac{Initial\ Investment}{Cash\ inflow\ per\ period} \tag{29}$$

The following formula to calculate the net present value NPV gathers the total net benefits of a project upon deducting the costs, thus turning the final value into dollar terms as presented below;

$$NPV = \sum_{t=1}^{T} \frac{C_t}{(1+r)^t} - C_o \tag{30}$$

where; Ct = net cash inflow during the period t, Co= total initial investment costs, r = discount rate, and t = number of time periods

The internal rate of return (IRR) is favorably used for comparing projects. It is simply defined as the interest rate which

the project is able to produce for the investors, and is calculated as the discount value that when applied in the NPV formula drives the NPV formula to zero. The formula for IRR is:

$$0 = P0 + \frac{P_1}{(1+IRR)} + \frac{P_2}{(1+IRR)^2} + \frac{P_3}{(1+IRR)^3} \\ ..... + \frac{P_n}{(1+IRR)^n} \quad (31)$$

where, P_0, P_1, . . . P_n equals the cash flows in periods *1, 2, . . .n*, respectively.

The levelized cost of electricity (LCOE) in electrical energy production is considered an essential tool to decide whether to move forward with a said project as it allows us to compare varied technologies of unequal life spans, project size, different capital cost, risk, return, and capacities. It is also described as the current value of the cost of the electrical energy generated over an assumed lifetime in cents/kW.hr. In other words, it is the cost of electricity needed when revenues equal costs. Thus, obtaining a RoC invested equivalent to rebate rate. The following formula exemplifies (LCOE) [29]:

$$LCOE = \frac{\sum_{t=0}^{t=n} \frac{I_t + O\&M_t - PTC_t - D_t + T_t + R_t}{(1+i)^t}}{C\sum_{t=0}^{t=n-1} \frac{P_t}{(1+i)^t}} \quad (32)$$

where *LCOE* is Generation Cost (cents/kWh), I_t is Investment made per year ($), $O\&M_t$ is Operating and maintenance per year ($), PCT_t is Production Tax Credit ($), D_t is Depreciation credit ($) , T_t is Tax levy ($), R_t is Land rent ($), P_t is Electrical generation capacity (Kwh) and *i* is the Discount rate fraction.

Table 3: Constant cost analysis, operation costs, and LEOC for both systems.

Item	Off- Grid	On-Grid
System Size (kW-DC)	$25.20	$9.13
1st-Year Production (kWh)	$17,986.36	$13,017.00
Annual Degradation	0.50%	0.50%
Cost ($/W)	$1.47	$1.57
Initial Rebate/Incentive	$2,343.05	$2,343.05
Total Investment cost	$36,930.00	$9,137.00
O&M Cost ($/kW)	$10.00	$15.00
O&M Escalator (%)	3.00%	3.00%
PPA Rate ($/kWh)	$0.13	$0.13
PPA Escalator	3.00%	3.00%
PPA/20 years	$0.1738	$0.1738
PPA/20 years	$0.1881	$0.1881
LCOE Outputs/20 years	$0.1205	$0.0631
LCOE Outputs/25 years	$0.1033	$0.0554
Discount rate	0.06	0.06
NPV	15097.56	17737.57
IRR	10.66%	25.36%

Table 3 exhibits the analysis for constant cost analysis, operation and maintenance costs, payback period, NPV, IRR, and LEOC for each item. Figure 10 below illustrates the cumulative cash flow for both systems. In general, the feasibility of installing a grid connected system is higher than standalone. Grid connected systems are usually cheaper with better efficiency rates considering the costs associated with the batteries and other standalone equipment which add to cost as well as maintenance. However, standalone systems provide a good alternative to supply remote areas or remote loads in which the investment in installing new electric network is quite higher compared to installing a standalone system. Choosing between a grid-connected and a standalone system is subject to case specifications, and is therefore deemed for the discretion of decision makers.

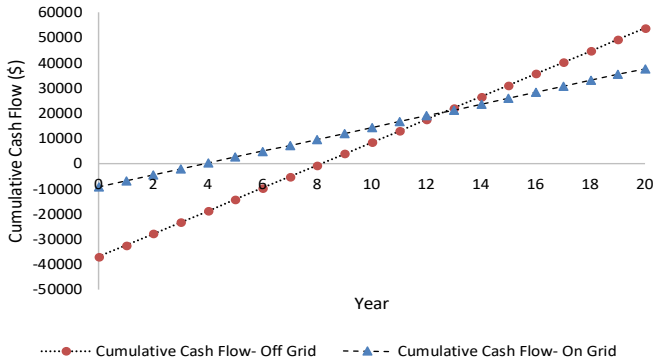


Figure 10. The cumulative cash flow for standalone and grid connected systems.

6. Conclusion

DG's flexibility, small size, minimum investment expenditure, and ecofriendly interests have the potential to improve distribution system performance. Moreover, DG also supplies increased load over time which in turn has a significant impact decreasing the needed investment. DGs have an advantage over central power plants since they can be mounted at or close to the load. The impact of distributed generation has multi-faceted effects on the system and its machinery including stability, reliability, steady-state procedure, power quality, dynamic procedure, and safety for consumers and power suppliers. Effects could be classified as either positive or negative, contingent upon several factors including the distribution system, distributed generators locations and sizing, and the load characteristics. Positive effects are usually referred to as "system support benefits," and entail: decrease in loss, enhanced reliability of the utility system, voltage support and enhanced power quality, transmission and distribution capacity release, deferrals of new or upgraded transmission and distribution infrastructure, quick and convenient installation on account of mass-produced components, reducing expenses by evading long distance high voltage transmission, and nature friendly where renewable sources are used.

Optimal allocation of distributed resources will in turn lead to realizing the potential positive impacts of distributed generation. Particularly, because it utilizes the actual function, limits and utilizes a suitable optimization algorithm. To that end, financial performance indicators (internal rate of return (IRR), net present value (NPV) and payback periods) have been assessed for both systems. To conclude, grid connected systems have the upper hand in terms of technical feasibility and economic viability compared to standalone system.

References

[1] R. Viral and D. K. Khatod, "An analytical approach for sizing and siting of DGs in balanced radial distribution networks for loss minimization,"

International Journal of Electrical Power & Energy Systems, vol. 67, pp. 191-201, 2015/05/01/ 2015.

[2] M. Al-Maghalseh, S. Odeh, and A. Saleh, "Optimal sizing and allocation of DGs for real power loss reduction and voltage profile improvement in radial LV networks," in *Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), 2017 14th International Conference on*, 2017, pp. 21-25: IEEE.

[3] O. Badran, S. Mekhilef, H. Mokhlis, and W. Dahalan, "Optimal reconfiguration of distribution system connected with distributed generations: A review of different methodologies," *Renewable and Sustainable Energy Reviews,* vol. 73, pp. 854-867, 2017/06/01/ 2017.

[4] T. Adefarati and R. C. Bansal, "Reliability assessment of distribution system with the integration of renewable distributed generation," *Applied Energy,* vol. 185, pp. 158-171, 2017/01/01/ 2017.

[5] C. Orozco-Henao, A. S. Bretas, R. Chouhy-Leborgne, A. R. Herrera-Orozco, and J. Marín-Quintero, "Active distribution network fault location methodology: A minimum fault reactance and Fibonacci search approach," *International Journal of Electrical Power & Energy Systems,* vol. 84, pp. 232-241, 2017/01/01/ 2017.

[6] M. R. Kaveh, R.-A. Hooshmand, and S. M. Madani, "Simultaneous optimization of re-phasing, reconfiguration and DG placement in distribution networks using BF-SD algorithm," *Applied Soft Computing,* vol. 62, pp. 1044-1055, 2018/01/01/ 2018.

[7] A. Soroudi, A. Rabiee, and A. Keane, "Distribution networks' energy losses versus hosting capacity of wind power in the presence of demand flexibility," *Renewable Energy,* vol. 102, pp. 316-325, 2017/03/01/ 2017.

[8] S. Pukhrem, M. Basu, M. F. Conlon, and K. Sunderland, "Enhanced Network Voltage Management Techniques Under the Proliferation of Rooftop Solar PV Installation in Low-Voltage Distribution Network," *IEEE Journal of Emerging and Selected Topics in Power Electronics,* vol. 5, no. 2, pp. 681-694, 2017.

[9] G. Pillai, G. Putrus, N. Pearsall, and T. Georgitsioti, "The effect of distribution network on the annual energy yield and economic performance of residential PV systems under high penetration," *Renewable Energy,* vol. 108, pp. 144-155, 2017/08/01/ 2017.

[10] R. Torquato, D. Salles, C. O. Pereira, P. C. M. Meira, and W. Freitas, "A Comprehensive Assessment of PV Hosting Capacity on Low-Voltage Distribution Systems," *IEEE Transactions on Power Delivery,* vol. 33, no. 2, pp. 1002-1012, 2018.

[11] V. Behravesh, R. Keypour, and A. A. Foroud, "Stochastic analysis of solar and wind hybrid rooftop generation systems and their impact on voltage behavior in low voltage distribution systems," *Solar Energy,* vol. 166, pp. 317-333, 2018/05/15/ 2018.

[12] E. McKenna, J. Pless, and S. J. Darby, "Solar photovoltaic self-consumption in the UK residential sector: New estimates from a smart grid demonstration project," *Energy Policy,* vol. 118, pp. 482-491, 2018/07/01/ 2018.

[13] F. M. Camilo, R. Castro, M. E. Almeida, and V. F. Pires, "Assessment of overvoltage mitigation techniques in low-voltage distribution networks with high penetration of photovoltaic microgeneration," *IET Renewable Power Generation*, vol. 12, no. 6, pp. 649-656Available: http://digital-library.theiet.org/content/journals/10.1049/iet-rpg.2017.0482

[14] M. Al-Maghalseh and W. Saleh, "Design and cost analysis of biogas based power plant: Jenin perspective," in *Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT), 2017 14th International Conference on*, 2017, pp. 31-35: IEEE.

[15] M. M. Al-Maghalseh, A. K. Abutemeha, and M. Iyadiyyeh, "Modeling a hybrid system for electrical generation and wastewater treatment using photovoltaic and fuel cells," (in English), *Desalination and Water Treatment,* vol. 100, pp. 250-257, 2017.

[16] M. M. A. Al-Maghalseh, "Evaluating the Reliability worth Indices of Electrical Medium Voltage Network: Case Study," *Procedia Computer Science,* vol. 130, pp. 744-752, 2018/01/01/ 2018.

[17] K. A. W. Horowitz, B. Palmintier, B. Mather, and P. Denholm, "Distribution system costs associated with the deployment of photovoltaic systems," *Renewable and Sustainable Energy Reviews,* vol. 90, pp. 420-433, 2018/07/01/ 2018.

[18] M. A. Omar and M. M. Mahmoud, "Grid connected PV- home systems in Palestine: A review on technical performance, effects and economic feasibility," *Renewable and Sustainable Energy Reviews,* vol. 82, pp. 2490-2497, 2018/02/01/ 2018.

[19] P. Chaudhary and M. Rizwan, "Voltage regulation mitigation techniques in distribution system with high PV penetration: A review," *Renewable and Sustainable Energy Reviews,* vol. 82, pp. 3279-3287, 2018/02/01/ 2018.

[20] D. S. Pillai and N. Rajasekar, "A comprehensive review on protection challenges and fault diagnosis in PV systems," *Renewable and Sustainable Energy Reviews,* vol. 91, pp. 18-40, 2018/08/01/ 2018.

[21] I. Ali, G. M. Shafiullah, and T. Urmee, "A preliminary feasibility of roof-mounted solar PV systems in the Maldives," *Renewable and Sustainable Energy Reviews,* vol. 83, pp. 18-32, 2018/03/01/ 2018.

[22] A. Zurita *et al.*, "State of the art and future prospects for solar PV development in Chile," *Renewable and Sustainable Energy Reviews,* vol. 92, pp. 701-727, 2018/09/01/ 2018.

[23] D. K. Khatod, V. Pant, and J. Sharma, "Evolutionary programming based optimal placement of renewable distributed generators," *IEEE Transactions on Power Systems,* vol. 28, no. 2, pp. 683-695, 2013.

[24] D. Q. Hung and N. Mithulananthan, "Multiple Distributed Generator Placement in Primary Distribution Networks for Loss Reduction," *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS,* vol. 60, no. 4, 2013.

[25] D. Q. Hung, N. Mithulananthan, and R. C. Bansal, "Analytical strategies for renewable distributed generation integration considering energy loss minimization," *Applied Energy,* vol. 105, pp. 75-85, 2013/05/01/ 2013.

[26] R. S. A. Abri, E. F. El-Saadany, and Y. M. Atwa, "Optimal Placement and Sizing Method to Improve the Voltage Stability Margin in a Distribution System Using Distributed Generation," *IEEE Transactions on Power Systems,* vol. 28, no. 1, pp. 326-334, 2013.

[27] Y. M. Atwa, E. F. El-Saadany, M. M. A. Salama, and R. Seethapathy, "Optimal Renewable Resources Mix for Distribution System Energy Loss Minimization," *IEEE Transactions on Power Systems,* vol. 25, no. 1, pp. 360-370, 2010.

[28] Y. M. Atwa and E. F. El-Saadany, "Probabilistic approach for optimal allocation of wind-based distributed generation in distribution systems," *IET Renewable Power Generation,* vol. 5, no. 1, pp. 79-88, 2011.

[29] M. M. Al-Maghalseh and E. M. Maharmeh, "Economic and Technical Analysis of Distributed Generation Connection: A Wind Farm Case Study," *Procedia Computer Science,* vol. 83, pp. 790-798, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 281-288 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Probabilistic Method for Anomalies Detection Based on the Analysis of Cyber Parameters in a Group of Mobile Robots

Elena Basan*, Alexander Basan, Oleg Makarevich

Department of Information Security Southern Federal University, Taganrog, 347922, Russia

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 26 September, 2018
Online: 18 November, 2018

Keywords:
Mobile robots
Abnormal behavior
Probabilistic methods
Attack
Detection

ABSTRACT

This article is devoted to the issues of ensuring the security of a group of mobile robots in the implementation of attacks aimed at the property of accessibility of information and the availability of network nodes. The article presents a method for detecting an abnormal behavior of a network node based on the analysis by the group members of the parameters: residual energy and network load. Analysis of the behavior of individual robots relative to general behavior using probabilistic methods avoids the problem of creating a reference distribution for describing the behavior of a node, as well as creating a signature database for detecting anomalies. The developed method demonstrates high detection rate of denial of service attack and distributed denial of service attack with the number of malicious nodes not exceeding or slightly exceeding the number of trusted nodes. It also provides detection of the Sybil attack.

1. Introduction

This paper is an extension of work originally presented in CyberC 2017, "9th International Conference on Cyber-enabled distributed computing and knowledge discovery" titled 'A Trust Evaluation Method for Active Attack Counteraction in Wireless Sensor Networks' [1]. Mobile robot networks are quite vulnerable to attacks both over the network and physical properties of nodes. The article [2] presented a threats model for the network of mobile robots. The authors also analyzed the attacks for a group of mobile robots. Based on the analysis, it was revealed that the main set of attacks that an attacker can implement for a group of mobile robots is denial of service (DoS), distributed denial-of-service (DDoS) attack, a man in the middle (MITM) and a Sybil attack, and exhaustion resources. In addition, there are a number of attacks aimed at the robot positioning system and on other elements of the sensor system, which are not considered in this study. The main purpose of this study is to detect these attacks with a minimum of resources of mobile robots. The offender, implementing an active attack, can influence any physical parameter of the mobile robot through a network or physical impact.

1.1. Maintaining the Integrity of the Specifications

The standard methods used in intrusion detection systems (IDS) may not always have a positive effect when anomalous behavior of mobile robot network nodes is detected. This is due to several factors: 1. IDS, as a rule, work with the TCP / IP protocol stack, which is not always applicable to mobile robots that transmit data over the radio channel and can use any radio modules and any proprietary protocols; 2. Signature analysis, which is often used in the IDS [3], may be ineffective when an intruder is detected for a group of robots. The behavior of the group robots can vary significantly depending on the task being performed, including the level of network activity, it is quite difficult to create a signature database for the behavior of nodes in the context of each individual task. 3. The computing power of mobile robots is much lower than for standard computer systems for which the IDS is developed [4]. In addition, as a rule, mobile robots can either not be equipped with an operating system or have a "cut-down" version of an operating system with limited capabilities [5]. If the first two problems can be solved by writing their own software for IDS, then the problem of limited energy resources (in the form of insufficiently capacious batteries) makes the use of standard IDS almost impossible for mobile robots [6]. In [7], the authors considered an attack detection system based on the decision tree using the C5.0 algorithm applied to a group of robotic vehicles. The advantage of the presented approach is that for detecting cyber-attacks, the authors, along with four features for analyzing the process of communication and information processing, called cyber input functions, use four parameters for analyzing the physical properties of the robot, which the authors call the physical characteristics of the input signal. Next, the authors conduct 5 types of destructive impact on the robot and get a set of rules for building a decision tree. The

* Elena Basan, E-mail: ele-barannik@yandex.ru

disadvantage of the approach is that in this paper, attacks on only one robot, and not a network of robots were considered. At the same time, the authors considered a limited set of attacks: denial-of-service attacks and attacks aimed at violating physical parameters. In addition, in such systems, there is a need to constantly add rules to detect new attacks. This system is aimed at ensuring the availability of the transmitted data. In [8] an intrusion detection system based on the signature analysis is considered. The authors conducted a series of experiments to create a standard template describing the normal behavior of the robot in the absence of any external influence, as well as random behavioral anomalies. Then a number of situations in which abnormal behaviors occurred caused by environmental conditions were simulated. A normal behavior pattern of the robot based on the collected data with the weighting coefficients calculated on the basis of the frequency of occurrence of a particular type of abnormal behavior. This approach demonstrates greater efficiency in detecting a malicious node than a simple signature analysis; however, there are some disadvantages:

- The need to constantly update the signature database to control data from the new sensors of the mobile robot.
- Conducting analysis of changes only physical parameters of the node and the absence of network analysis of data.

The article [9] considers the system for detecting attacks on unmanned aerial vehicles. The development of this system used the approach based on the creation of a signature database. The system works as follows. Each node of the network has a monitor node, which may be a neighboring unmanned device that fixes the behavior of the trusted node and writes it to the matrix. The monitor node constantly monitors the behavior of the ward node and presents it with estimates. These estimates depend on how much the behavior of the ward deviates from the normal pattern of behavior. Then a rule database is created and the behavior of the node is evaluated. At the same time, the assessment is made on 7 parameters. The authors claim that their system is adaptive and demonstrates a low level of errors of the 1st and 2nd kind when detecting attacks. The disadvantage of the system is the need to constantly monitor the nodes one by one and analyze their behavior, which involves the computational load and network bandwidth.

The article [10] considers the system for detecting the abnormal behavior of robots of the Internet-robots network. The peculiarity of this system is that it has two subsystems. One is a group of robots that collect data using a sensor system and transmit it to the central node that is connected to an external mobile network. The second is a mobile network, where the following modules are available: a data acquisition module, an anomaly classification module, a control command module. The disadvantage of this system is that it is completely centralized; robots do not communicate with each other and act only through an intermediary. Anomaly detection occurs via classifier, which is trained by using training samples preformed. Thus, as a result of studying the works devoted to the topic of detecting attacks on robots, there are three main drawbacks in the existing approaches:

- Most systems based on signature analysis, either on a rules-based system. In this regard, there are the following limitations: the difficulty of detecting new attacks that are not related to the fixed patterns of the attacker's behavior, as well as the need to keep the database of rules or sets of signatures up-to-date.
- Systems based on fully distributed detection methods require additional energy costs, computational power costs from nodes and increase bandwidth. In addition, if the distributed system is used in conjunction with the signature analysis, the information about the abnormal behavior must be constantly updated, which uses the already limited resources of the robot's memory.
- When using centralized methods, a node that performs basic functions for detecting abnormal behavior is a vulnerability of the system.

In this article, a method for detecting an abnormal behavior of an attacker or several intruders within a group of mobile robots based on probabilistic methods is being developed [11]. The main difference of this method is that it does not require the creation of a standard probability distribution, like other probabilistic methods. The absence of the need to build a reference distribution is due to the fact that the current indications of the node group are taken to reveal the anomalous behavior, then the normal distribution function is constructed and the confidence interval of values is calculated. To estimate the behavior of the Ni node, the probability of the current node indicators entering the confidence interval is calculated, based on the indices of all nodes of the group. Thus, it becomes possible to estimate the probability of the node deflection behavior of the overall behavior of a group of nodes.

2. Method for detecting abnormal behavior

The peculiarity of the proposed method for a group of robots is that for the formation of the normal distribution function it is necessary to obtain data from several nodes performing a similar set of actions. To more accurately determine the degree of deviation of the current indications of a node from a group of nodes, it is necessary that the indicators of a group of nodes are in the same range. In the case of a group of mobile robots that exchange information in one task, this method will work most efficiently. An attacker can affect both the network connection between nodes and the physical parameters of the network node. Table 1 shows the parameters that can be affected by the attacker and the attack by which he can do it.

The parameter packets with data - here it is understood the fact of transfer or redirection of the packet, that is, the availability of the transmitted information is estimated. If there is any impact on the network from the attacker, then there may be situations when packets are discarded, duplicated, etc.

Table 1. The correlation of network parameters and the attacks affecting them

Parameters (indicators of anomalies)	Attacks
Data packets	Black Hole, Gray Hole, False Redirection, Denial of Service, Packet Delay
Remaining battery power	Denial of service, depletion of resources
Network load	Denial of service, resource depletion, the Sybil attack, Flood-attack , Wormhole
Package Integrity	Modification, substitution messages, Man in the middle

The battery charge parameter is the current consumption of the battery (or power consumption), as well as the remaining energy

reserve in the battery pack, which allows the device to function in the network [12].

The Network Load - the total number of packets transmitted on the network. Either the number of packets transmitted through one of the nodes of the network [13].

When detecting attacks such as denial of service and attacks aimed at depleting resources, it is necessary to select those parameters that will be evaluated according to the claimed method. Thus, in the case of a denial of service attack changing network load for a malicious node and network load for the victim. Therefore, it is advisable to estimate the network load, which is the total number of received, sent, and redirected packets of the network node. When an attacker implements an attack aimed at depleting the resources of the node, there will be a sharp decrease in the residual energy level of the node. In addition, an attacker can have superiority in the reserves of energy resources. In the implementation of the Sibyl attack or attack redirecting the impact on themselves, the main purpose of the attacker is to change the processes and routes of data exchange in the network. In other words, an attacker achieves such a situation that all or most of the traffic of neighboring nodes passes through him. Further, the attacker can simply drop received packets, or send them to the wrong nodes. The attacker can achieve this situation in various ways, in this case it is important that the level of incoming traffic will be much higher than that of other nodes of the network. Therefore, it is also necessary to consider the network boot parameter to detect this attack. In the previous work of the authors [14], in addition to the parameters, network loading and residual energy, the parameter of the number of discarded packets P1 was considered. In this study, it will not be considered. Thus, consider two parameters: the network load P2 and the residual energy P3. Changing these parameters affects the state of both the nodes of the network and the entire group of robots in general. The state of the nodes of the network can be described as follows: S1 - the state when the node is not subject to attack and does not conduct the attack itself, i.e. is authentic at the current time; S2 - the state when the behavior of the mobile robot deviated from the behavior of the greater part of the robot group can be observed provided that the node became the victim of the attack, i.e. the node is undefined; S3 - when the behavior of the mobile robot is significantly different from the nodes of the group, i.e. most likely the site is malicious. Figure 1 shows the transition graph from one state to another, and also reflects the effect of parameters and attributes on each other.

The following attributes of the node affect the parameter $P2 = L$: A_{21} - the total number of packets sent by the node containing data. In this model, data transfer uses the UDP protocol and the CBR traffic type. $A_{21} = scbr$; A_{22} is the total number of management packs, or beacons, for testing connections. In this model, packets sent via the ARP protocol act in this role. $A_{22} = s_{arp}$; A_{23} is the total number of packets sent over the routing protocol. This model uses the AODV routing protocol. $A_{23} = s_{aodv}$; A_{24} is the total number of received CBR packets. $A_{24} = r_{cbr}$; A_{25} - the total number of ARP packets received. $A_{25} = r_{arp}$; A_{26} is the total number of received AODV packets. $A_{26} = r_{aodv}$; A_{27} is the total number of dropped CBR packets. $A_{27} = d_{cbr}$; A_{28} is the total number of dropped ARP packets. $A_{28} = d_{arp}$; A_{29} is the total number of discarded AODV packets. $A_{29} = d_{aodv}$.

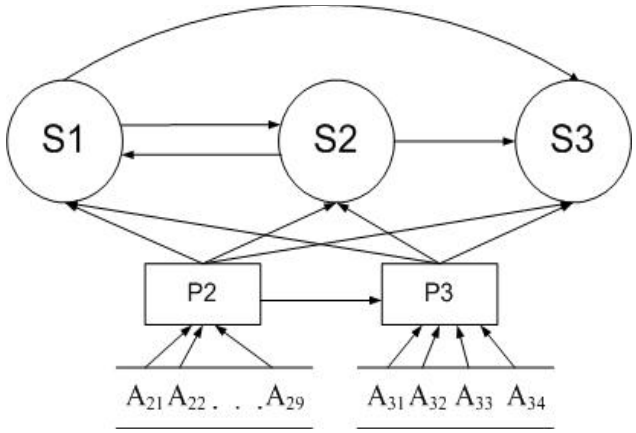


Figure.1. Graph of state of nodes in groups of mobile robots

Thus, the parameter L can be represented by the following equation:

$$L = A_{21} + A_{22} + A_{23} + A_{24} + A_{25} + A_{26} + A_{27} + A_{28} + A_{29} \quad (1)$$

Parameter $P_3 = e$ can be characterized by a finite set of attributes A_{3j}. The following attributes affect the amount of residual energy of the node, but apart from the attributes described below, the amount of residual energy is affected by the node's load, i.e. parameter $P2$: A_{31} is the initial energy reserve of the network node. $A_{31} = initialEnergy$; A_{32} is the power of the transmitted signal. $A_{32} = rxPower$; A_{33} - signal reception power, $A_{33} = txPower$; A_{34} - speed of moving the node. $A_{34} = speed$.

The residual energy is calculated by reducing the level of initial energy A_{31} for each transmitted and each received packet per unit time:

$$e_{tx} = A_{31} - (A_{33} * txtime)$$
$$e_{rx} = A_{31} - (A_{32} * rcvtime) \quad (2)$$

where e_{tx} and e_{rx} this is the level of residual energy after receiving the packet and after the transmission of the first packet; $rcvtime$ - time of packet transmission; $txtime$ - time of reception of a package. A formal description of the method is presented in Section 3, which provides an algorithm implemented in a simulation system for conducting an experimental study.

3. Implementation of the method for detecting abnormal behavior

The simulation of the developed method was carried out in the simulator NS-2.35. The procedure for detecting abnormal behavior and outputting results is called regularly at regular intervals. The start and end time of this process is set by the user in the script using a special command. The command handler plans to start a special timer for the start time of the process, and the timer and end time are written to the timer parameters [15]. To account for the parameters e and L of the mobile robot in the assembly model was added a special object that counts the number of packets transmitted / sent / forwarded node and the residual energy of the node [16]. The procedure for calculating trust works according to the following algorithm representing at the table 2.

Table 2: Calculation of the trust level for nodes in the group of mobile robots

No.	Name of equation	Equation	Description
1	**The calculation of the confidence interval boundaries**		
1.1	Variance for L parameter	$D_{Li} = \left(\sum_{i}^{N} (L_i - \overline{L})^2 \right) / n$	D_{Li}, - variance of the parameters L calculated for the group of nodes in the current time interval
1.2	Variance for e parameter	$D_{ei} = \left(\sum_{i}^{N} (e_i - \overline{e})^2 \right) / n$	D_{ei} - variance of the parameters e calculated for the group of nodes in the current time interval
1.3	The standard deviation for L parameter	$\sigma_{Li} = \sqrt{D}_{Li}$	σ_{Li} - is the standard deviation of the parameter L which calculated for the group of nodes in the current time interval.
1.4	The standard deviation for e parameter	$\sigma_{ei} = \sqrt{D}_{ei}$	σ_{ei} - is the standard deviation of the parameter e which calculated for the group of nodes in the current time interval.
1.5	Argument of the Laplace function - t	$\Phi(t) = \frac{\alpha}{2}$	$\Phi(t)$ - is the Laplace function; α is a given reliability, in this study the value of the coefficient is equal to $\alpha = 0.98$, so the argument $t = 2.34$;
1.6	The limits of the confidence interval for the e parameter	$a_{\min} = \overline{e} - t \cdot \sigma_e / \sqrt{n}$, $a_{max} = A_{31}; a_{min} < a_{max}$	The upper bound of the confidence interval of the parameter e is always equal to the maximum permissible energy value, that is, $a_{max} = initialEnergy$. This is due to the fact that nodes can migrate from one group to another; new nodes may appear with a residual energy value equal to the initial value. $a_{\min}$ - lower bound of confidence interval. $t * \sigma / \sqrt{n}$ - is the accuracy of the estimation. n - Total number of nodes.
1.7	The limits of the confidence interval for the L parameter	$b_{min} = L_{min}$, $b_{max} = \overline{L} + t \cdot \sigma_L / \sqrt{n}$	The lower bound for the parameter L is equal to the minimum required number of packets passed through the node in one time interval Lmin. These measures are taken because the mobile robot can exhibit selfish behavior, that is, refuse to participate in the network to save energy, which can artificially "understate" the boundaries of the interval. b_{max} - the upper bound of confidence interval for L parameter
2	**Determination of the probability of anomalous behavior of the mobile robot on the basis of the calculated confidence intervals.**		In order to calculate the mean square deviation and mathematical expectation, it is necessary to shorten the interval for which the value is calculated and take into account only the node parameters in the previous time interval L_{i-1}, e_{i-1} and L_i, e_i the node parameters for the current interval. Note: If you take the parameter values over the entire time interval, the standard deviation is too large, due to the large difference between the start and end values.
2.1	The mathematical expectation for the e parameter	$\overline{e_в} = (e_{i-1} + e_i) / 2$	$\overline{e_в}$ - mathematical expectation of the values of the e parameter for the sampling interval, which calculated for individual node
2.2	The mathematical expectation for the L parameter	$\overline{L_в} = (L_{i-1} + L_i) / 2$	$\overline{L_в}$ - mathematical expectation of the values of the L parameter for the sampling interval, which calculated for individual node
2.3	The variance for the e parameter	$D_{e_в} = \left(\sum_{i}^{N} (e_i - \overline{e_в})^2 \right) / n$	$D_{e_в}$ - variance for the sampling interval for e, which calculated for individual node.
2.4	The variance for the L parameter	$D_{L_в} = \left(\sum_{i}^{N} (L_i - \overline{L_в})^2 \right) / n$	$D_{L_в}$ - variance for the sampling interval for L, which calculated for individual node.
2.5	The standard deviation for the e parameter	$\sigma_{e_в} = \sqrt{D}_{e_в}$	$\sigma_{e_в}$ - the standard deviation for the sampling interval for the residual energy, which calculated for individual node.
2.6	The standard deviation for the L parameter	$\sigma_{L_в} = \sqrt{D}_{L_в}$	$\sigma_{L_в}$ - the standard deviation for the sampling interval for the L parameter, which calculated for individual node.

2.7	The probability of the value for parameter e falling into the confidence interval	$P_e(a_{\min} < e_i < a_{\max}) = = \Phi\left(\frac{a_{\max} - \overline{e_e}}{\sigma_{e_e}}\right) - \Phi\left(\frac{a_{\min} - \overline{e_e}}{\sigma_{e_e}}\right)$	P_e,- the probability of deviations the network load from confidence interval, which calculated for individual node. Φ is a Laplace function.
2.8	The probability of the value for parameter L falling into the confidence interval	$P_L(b_{\min} < L_i < b_{\max}) = = \Phi\left(\frac{b_{\max} - \overline{L_e}}{\sigma_{L_e}}\right) - \Phi\left(\frac{b_{\min} - \overline{L_e}}{\sigma_{Le}}\right)$	P_L - the probability of deviations the residual energy from confidence interval, which calculated for individual node.
2.9	The resulting probability value that the node is trusted	$P_{sum} = P_e * P_L$	To obtain the resulting probability value, it is necessary to use a combination of the values of P_e,, P_L of the direct value of trust P_{sum} in [17], an algorithm for combining confidence values using the Bayes theorem is presented.
3.	Deciding on the degree of trust in the node	$P_{sum} > 0,5;$ $P_{sum} = 0,5;$ $P_{sum} < 0,5.$	Assume threshold probability that the node is abnormal equal to 0.5. When the node reaches a value of 0.5, it is necessary to reduce its residual energy level by half, then the node is considered in an undefined state . Further, if the value of the confidence level reaches the level of 0.4, then it is necessary to consider the node malicious and reduce its energy level to zero, thus, the node is excluded from the network [18].

4. Experimental study, evaluation of the effectiveness of the developed method.

The model of a robot group in the simulation environment NS-2.35 was developed. Robots communicate with each other via wireless communication and use the TCP / IP protocol stack to transfer information. In particular, the UDP protocol is used for data transmission at the transport level, the ARP protocol is used to transmit control commands at the data link layer, the AODV protocol is used for routing the packets [19]. Figure 2 shows a group of mobile robots in the modeling system, which includes 10 nodes. Of these, one N4 node is a base station or a central server. The node N0 is the group leader and performs the functions of gathering information from the other robots and redirects it to the central server. The nodes of the group exchange information with each other and with the group leader [20]. In this case, nodes N6, N7, N8, N9 will conduct a DDoS attack starting from 50 seconds of network operation.

4.1. Implementation and detection of a DoS attack.

Conducting denial of service attack, the attacker creates a situation where the network node becomes unavailable to other nodes and cannot respond to their requests and work normally. An attack aimed at depleting resources, as a rule, creates such conditions for a node that it begins to lose more energy than in the absence of an attack. These attacks are interrelated. In fact, the goal of a DoS attack can be to completely disable a node by exhausting the node's resources. The developed model of a group of mobile robots is assumed that the nodes spend energy on the transmission and reception of packets. Therefore, an attacker "forces" trusted nodes to spend more energy than when working in normal mode, sending a large number of packets to the network.

Three types of situations were simulated. In the first case, an estimate was made of the energy consumed by network nodes in the absence of an attack. In Figure 3, this situation is represented by a blue chart marked with rhombuses.

In the second case, an attack was conducted on the network, while the traffic of the malicious node is $I_t \le I_m \le 2I_t$, where I_m is the traffic of the malicious node, It is the traffic of the authentic node. The third graph represents a situation where an attack is carried out intensively and $I_m > 2\ I_t$.

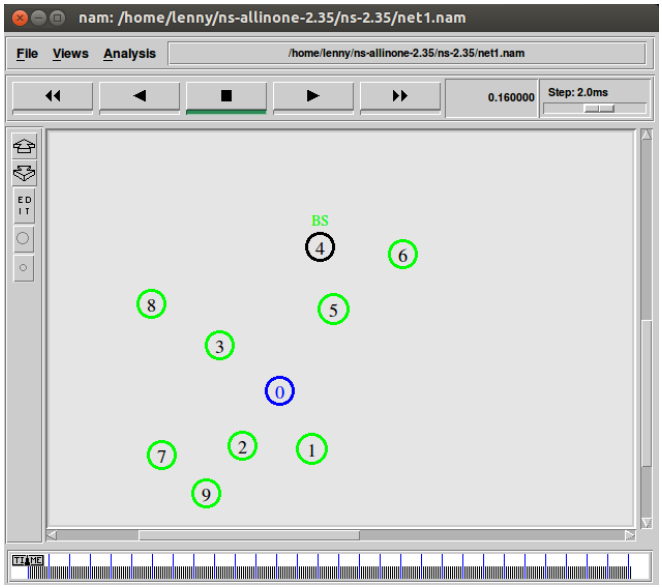


Figure 2. Group of robots in the simulation system NS-2.35

Figure 3 shows that during a non-intense attack, the energy level of the nodes will remain almost the same as for the case when the attack is not carried out. That is, in this case, the attack can be considered ineffective. The graph showing the change in the energy level during an intense attack shows a sharp drop in the energy level, which confirms the effectiveness of the attack. At the same time, the load of the attacker's node is more than twice the workload of authentic nodes. In this case, the developed method for detecting abnormal behavior allows us to identify a malicious node.

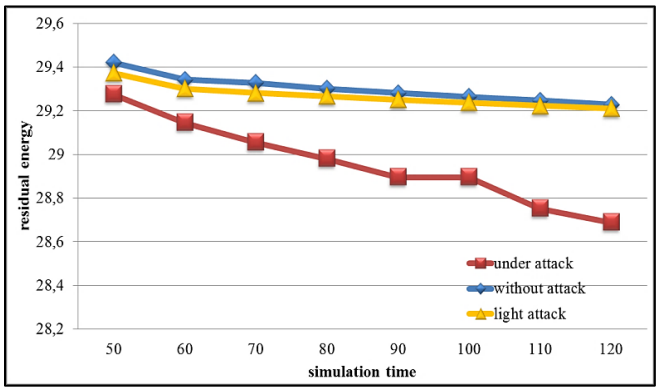


Figure 3. Change in the level of residual energy, depending on traffic intensity of network nodes.

The proposed method allows in a few seconds to detect a DoS attack, if it has a significant impact on the resources of network nodes and helps to increase the level of traffic. In Figure 4, a graph showing the level of detection of an attack, given that an attacker starts an attack after the 50th second of simulation, we can say that the attack is almost immediately detected. Since on an interval of time between 50-60th seconds the malicious node has a hit level in the confidence interval of 0.4 and is already blocked by the system for 60 seconds.

4.2. Implementation and detection of DDoS attacks.

The detection of a distributed denial of service attack is more difficult. This is due to the fact that when the number of malicious nodes prevails over the number of authentic nodes, the boundaries of the confidence interval are significantly expanded. Especially if malicious nodes conduct an attack with varying intensity. Nevertheless, the developed method is quite effective in detecting this attack. When the ratio of malicious nodes to trusted hosts is 4 to 5, the method allows to immediately detect all malicious nodes and block them already in the second time interval.

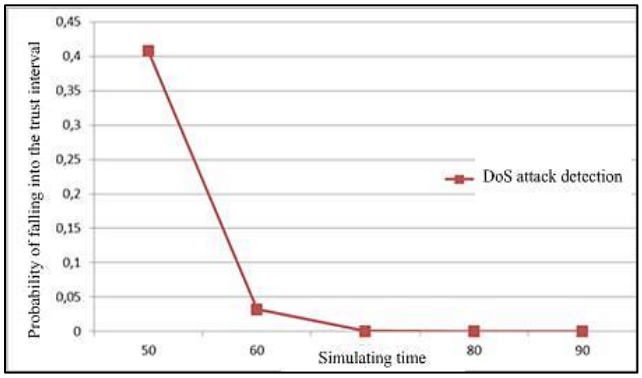
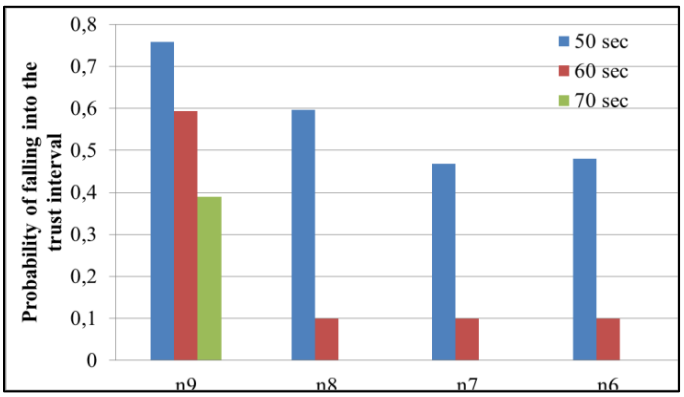


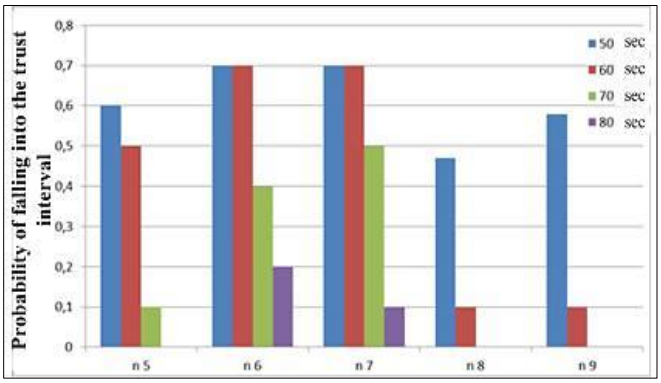
Figure 4. The probability of hit by the load and residual energy values of the malicious node in the trust interval

Figure 5 (a) presents a histogram showing the level of hit of current indicators e and L of malicious nodes in the confidence interval. When malicious and trusted hosts are in an equal ratio of 5 to 5, the quality of detection becomes worse.

Figure 5 (b) shows a histogram showing the detection level of malicious nodes. N9 and N8 nodes were also detected in the second interval, nodes N5 and N6 were detected in the third interval and node N7 in the fourth time interval, starting from the moment when the attack began. In general it can be said that the detection rate of 100%, but the rate of detection decreased.



(a)



(b)

Figure. 5. The level of detection of an attacker in a distributed denial of service attack for (a) four malicious hosts (b) for five malicious nodes and five authentic hosts

When the number of malicious nodes exceeds the number of trusted in the ratio of 6 malicious to 5 authentic, the detection level is 83%, i.e. one malicious node remains undetected. Figure 6 shows a histogram of the level of hit of current values of malicious nodes in the confidence interval. In this case, three nodes: N8, N9, N5 - are detected in the second time interval. One node N6 in the third interval and one node N10 in the 4th interval, only node n7 remains undetected on the fourth interval, most likely if the attack continues at the same rate, then this node will be detected on the 5th interval. But this time is high enough to detect an attack [21]. Nevertheless, the developed method shows a sufficiently high speed of detection of attack, even if the number of malicious nodes is more than the number of trusted ones.

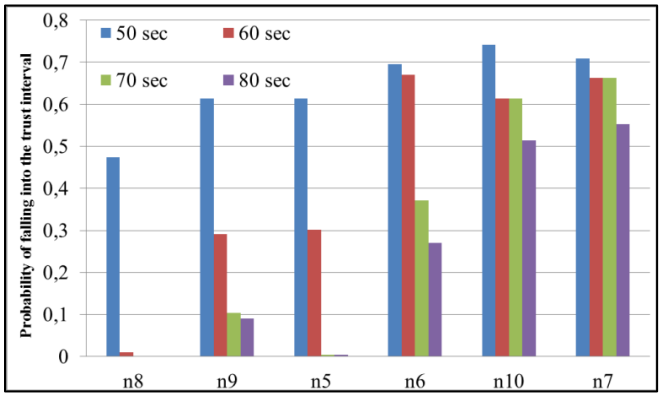


Figure. 6. The detection level of nodes with distributed attack denial of service for 6 malicious nodes and 5 trusted ones

4.3. Implementation and detection of the Sibyl attack.

The Sibyl attack is that an attacker is represented by several network nodes and tries to redirect most of the traffic to itself [22]. At the same time, it can make a destructive impact on the network by discarding messages, redirecting them to the wrong nodes, violating the routing scheme, or can passively listen for traffic.

At the same time to detect an attack, when an attacker does not have a destructive effect is quite difficult. In the works of the authors [23], as a rule, there are methods using hard protection: password protection, cryptographic protection, as well as signature analysis and group detection. These methods are used in networks MANET, IoF, P2P [24], which are not so much limited in resources as groups of mobile robots.

The developed method for detecting abnormal behavior shows the effectiveness of detection of the Sibyl attack, even if the attacker redirects the traffic to himself and does not take any further action. In this case, detection is possible by changing the load level of nodes that conduct an attack on neighboring nodes. In addition, the level of residual energy of the attacking nodes is significantly reduced. To assess the method, malicious N7-N11 nodes were added to the robot group in the NS-2.35 simulation system, which are called (Sybil1-Sybil5). Figure 7 shows the topology of the network, taking into account malicious nodes. The figure shows that the number of malicious nodes and the number of trusted ones, excluding the base station (BS) and the group leader (GL), corresponds to half of the network nodes.

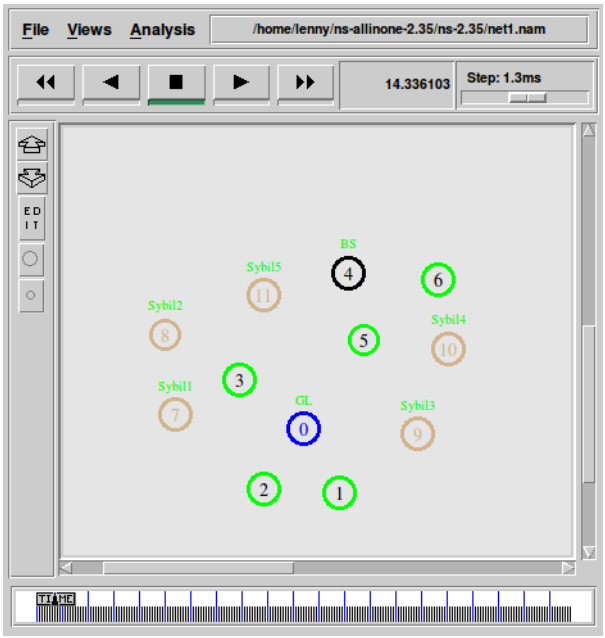


Figure. 7. The network topology for the implementation of the Sybil attack.

N7-N11 nodes redirect packets from neighboring mobile robots to themselves starting from 50 seconds, thus disrupting the network operation scheme. Initially, mobile robots will send packets to the group leader in a predetermined pattern; the leader of the base station sends packets. Thus, in the first 10 seconds of the attack, the method allows you to identify 2 malicious nodes N10 and N11. This is due to the fact that these nodes redirect more traffic to themselves. Further, starting from the 60th second, the detected nodes are blocked and nodes N7 and N8 are detected at the 70th second. The most difficult for detection was the node N9, this is due to its relatively low activity for redirecting traffic, at the time of detection the level of congestion of this node is less than twice the level of congestion of other nodes. Figure 8 shows a histogram representing the detection level of malicious nodes.

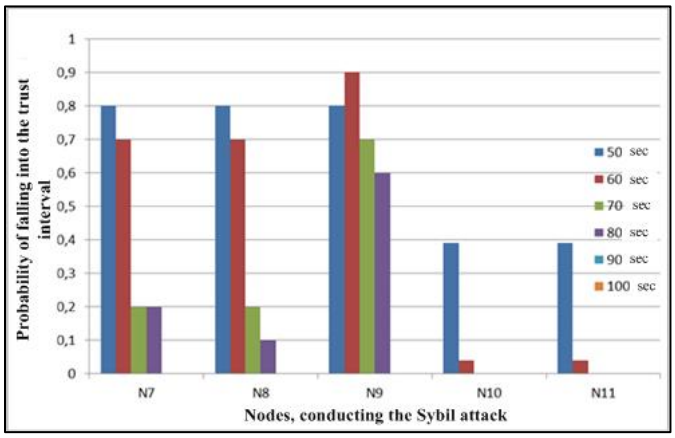


Figure. 8. detection level nodes conductive the Sybil attack

5. Conclusion

The issues related to the security of mobile robots and, in particular, group management of mobile robots, are currently being addressed by a limited number of scientists and institutions. Nevertheless, the subject of research is quite relevant, in connection with the widespread use of robotic systems. The developed method is a versatile tool for detecting anomalous behavior for a group of nodes, when it is possible to conduct an analysis of the behavior of most nodes and identify single or mass deviations from general behavior. Due to this, it is possible to increase the number of analyzed parameters for expanding the range of attacks. The method demonstrates a sufficiently high level of detection, namely it detects all malicious nodes within 30-40 seconds. Limitations of the method consist in the number of malicious nodes that conduct an attack on the network; their number should not exceed 60%, compared to the number of trusted ones. And also this method is applicable to a group of nodes that perform a similar task. In contrast to existing methods of detecting abnormal behavior, where one has to be comprehensive signature database or rules databases, store them, and then update the developed method makes it possible to detect anomalies in the current time, and depending on the current situation. This advantage is quite important because mobile robots can be used in different environments and for various tasks. In this case, in addition to networking protocols between mobile robots can change and environmental conditions, which in turn also influence the occurrence of anomalies and errors.

This method takes into account the threshold values, that is, the permissible level of anomalies, which reduces the number of false positives. In this case, the developed method allows to reduce the load on the mobile robot, it does not need to constantly exchange messages, monitor neighboring nodes and update, store the signature database. In the future study, it is proposed to add the node mobility parameter, to estimate the coordinates and speed of its movement. Evaluation of these parameters will allow to detect attacks such as interception by management, when an attacker, captures a trusted node and manages it independently.

Estimation of position and speed of movement will detect abnormal behavior nodes.

Acknowledgment

The work was supported by the Ministry of Education and Science of the Russian Federation (Initiative Science Projects No. 2.6244.2017 / 8.9).

References

[1] A. Basan, E. Basan, O. Makarevich. “A Trust Evaluation Method for Active Attack Counteraction in Wireless Sensor Networks”. in 9th International Conference on Cyber-enabled distributed computing and knowledge discovery. Nanjing, China.2017.P. 369-372. DOI: 10.1109/CyberC.2017.14.

[2] Basan A.S., Basan E.S. “Threat model for mobile robot group management systems”. System synthesis and applied synergetics. Collection of proceedings of the VIII All-Russian Scientific Conference. - Sistemnyy sintez i prikladnaya sinergetika. Sbornik nauchnykh trudov VIII Vserossiyskoy nauchnoy konferentsii. Rostov-on-Don: Southern Federal University. 2017. C. 205-212. - (In Russ.)

[3] H.S. Kim, S.W. Lee. “Enhanced novel access control protocol over wireless sensor networks”. IEEE Transactions on Consumer Electronics. 2009. № 55 (2). pp. 492 - 498.

[4] Branitsky A.A., Kotenko I.V. “Analysis and classification of methods for detecting network attacks”, Information security. Proceedings of SPIIRAN - Informatsionnaya bezopasnost'. Trudy SPIIRAN.2016. №2(45). pp.207-244.(In Russ.)

[5] Petrovsky O. “Attack on the drones”. Virus bulletin conference. 2015. pp. 16-24.

[6] Garber L. Robot OS: “A New Day for Robot Design”. Computer. № 46 (12). 2013. pp. 16-20.

[7] Kozhemyakin I.V., Putintsev I.A., Semenov N.N., Chemodanov M.N. “Development of an underwater robotic complex, using open simulation tools, supplemented with a model of hydroacoustic interaction”. News of SFedU. Technical science. Section II. Marine robotics. - Izvestiya YUFU. Tekhnicheskiye nauki. Razdel II. Morskaya robototekhnika. 2016. № 1 (174). pp.88-99.(In Russ.)

[8] Vuong T. P., Loukas G., Gan D., Bezemskij A. “Decision tree-based detection of denial of service and command injection attacks on robotic vehicles”. 2015 IEEE International Workshop on Information Forensics and Security (WIFS) 2015. pp. 1-6.

[9] Bezemskij A., Loukas G., Richard J. Gan D. “Behavior-based anomaly detection of cyber-physical attacks on a robotic vehicle”. 15th International Conference on Ubiquitous Computing and Communications and 2016 8th International Symposium on Cyberspace and Security. 2016. pp. 61-68.

[10] Mitchell R., Chen I.R. “Adaptive Intrusion Detection of Malicious Unmanned Air Vehicles Using Behavior Rule Specifications”. IEEE Transactions on Systems, Man, and Cybernetics: Systems. №44 (5). 2014. pp. 593 -599

[11] Monshizadeh M., Khatri V., Kantola R., Yan Z. “An Orchestrated Security Platform for Internet of Robots”. Springer International International Conference on Green, Pervasive, and Cloud Computing. 2017. pp. 298– 312.

[12] Basan A.S., Basan E.S., Makarevich O.B. “Method of counteracting active attacks of an attacker in wireless sensor networks”. News of SFedU. Technical science. №5 (190).2017. pp. 16-25. (In Russ.)

[13] Basan A., Basan E., Makarevich O. “Methodology of Countering Attacks for Wireless Sensor Networks Based on Trust”. 8th International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC).Chengdu.2016. pp.409 – 412.

[14] Schoch E., Feiri M., Kargl F., Weber M. “Simulation of Ad Hoc Networks: ns-2 compared to JiST/SWANS SIMUTools”. First International Conference on Simulation Tools and Techniques for Communications, Networks and Systems. 2008. pp. 34 - 41.

[15] He W., Yang S., Teng D., Hu Y. “A Link Level Load-Aware Queue Scheduling algorithm on MAC layer for wireless mesh networks”. International Conference on Wireless Communications & Signal Processing.2009. Nanjing, China.pp.1-16

[16] Basan A.S., Basan E.S., Makarevich O.B. “Development of the Hierarchal Trust management System for Mobile Cluster–based Wireless Sensor Network”. Proceeding SIN '16 Proceedings of the 9th International Conference on Security of Information and Networks. 2016. pp. 116-122.

[17] Mohammad Momani. Bayesian Fusion Algorithm for Inferring Trust in Wireless Sensor Networks // Journal of Networks 5(7). 2010; C. 815-822. DOI: 10.4304/jnw.5.7.815-822

[18] Abramov E.S., Basan E.S. “Development of a model of a protected cluster wireless sensor network”. News of SFedU. Technical science. - Izvestiya YUFU. Tekhnicheskiye nauki.2013. № 12(149). pp. 48-56. (In Russ.)

[19] Ferronato J. J. Sandini Trentin M. A. “Analysis of Routing Protocols OLSR, AODV and ZRP in Real Urban Vehicular Scenario with Density Variation”. IEEE Latin America Transactions. 2017. № 15 (9). pp. 1727 – 1734.

[20] Pshikhopov V.Kh., Soloviev V.V., Titov A.E., Finaev V.I., Shapovalov I.O. “Group management of mobile objects in uncertain environments”. Ed. V.H. Pshihopova. M.: FIZMATLIT. - Pod red. V.KH. Pshikhopova. M.: FIZMATLIюT 2015. – pp. 233-270. (In Russ.)

[21] Sargeant I., Tomlinson A. “Maliciously Manipulating a Robotic Swarm”. Int'l Conf. Embedded Systems, Cyber-physical Systems, & Applications. ESCS'16. 2016. pp. 122- 128.

[22] Abbas S., Merabti M., D. Llewellyn-Jones, K. Kifayat. “Lightweight Sybil Attack Detection in MANETs”. IEEE system journal, №. 7, (2). 2013. pp 236-248.

[23] Patel S.T., Mistry N.H. “A review: Sybil attack detection techniques in WSN”. 4th International Conference on Electronics and Communication Systems (ICECS). 2017. pp. 184 – 188.

[24] Wang G., Musau F., Guo S., Abdullahi M. B. “Neighbor Similarity Trust against Sybil Attack in P2P E-Commerce”. IEEE transactions on parallel and distributed systems. № 26 (3). 2015. pp. 824-833.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2, 283-308 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Design and Implementation of Closed-loop PI Control Strategies in Real-time MATLAB Simulation Environment for Nonlinear and Linear ARMAX Models of HVAC Centrifugal Chiller Control Systems

Nicolae Tudoroiu[*,1], Mohammed Zaheeruddin[1], Songchun Li[1], Elena-Roxana Tudoroiu[2]

[1]*Faculty of Engineering and Computer Science, Building Civil and Environmental Engineering, Concordia University, H3G 1M8, Canada*

[2]*Faculty of Sciences, Mathematics and Informatics, University of Petrosani, 332006, Romania*

A R T I C L E I N F O

Article history:
Received: 22 February, 2018
Accepted: 04 April, 2018
Online: 25 April, 2018

Keywords:
Centrifugal chiller
ARMAX and ARX models
PI control
PI-plus control
Parameters Estimation

A B S T R A C T

The objective of this paper is to investigate three different approaches of modeling, design and discrete-time implementation of PI closed-loop control strategies in SIMULINK simulation environment, applied to a centrifugal chiller system. Centrifugal chillers are widely used in large building HVAC systems. The system consists of an evaporator, a condenser, a centrifugal compressor and an expansion valve. The overall system is an interconnection of two main control loops, namely the chilled water temperature inside the evaporator, and the refrigerant liquid level control in condenser. The centrifugal chiller dynamics model in a discrete-time state-space representation is of high complexity in terms of dimension and encountered nonlinearities. For simulation purpose the centrifugal chiller model is simplified by using different approaches, especially the development of linear polynomials ARMAX and ARX models. The aim to build linear ARMAX models for centrifugal chiller is to simplify considerable the control design strategies that are investigated in this research paper. The novelty of this research is a new controller design approach, more precisely an improved version of proportional - integral control, the so called Proportional-Integral-Plus control for systems with time delay, based on linear ARMAX models. It is conceived within the context of non-minimum state space control system that "seems to be the natural description of a discrete-time transfer function, since its dimension is dictated by the complete structure of the model". The effectiveness of this new controller design, its implementation simplicity, convergence speed and robustness are proved in the last section of the paper.

1. Introduction

Nowadays significant amount of energy is consumed in commercial or residential buildings by the heating, ventilation and air conditioning (HVAC) systems to provide thermally comfortable indoor environment. Consequently, improving their efficiency becomes a critical issue for energy and environmental sustainability. In most commercial buildings, chilled-water from centrifugal chillers is supplied to air-handling units to meet the cooling needs of the building. The overall system as shown in Figure 1 consists of two water circuits: in one circuit water is circulated through the evaporator to produce chilled water and the second water circuit is used to reject heat to outdoors through a cooling tower [1-7]. During the last two decades, for a wide variety of HVAC control systems applications, the centrifugal chillers have become the most widely used due to their high capacity, high reliability, and low maintenance [7]. Moreover, among the major devices in chilled-water systems, the centrifugal chiller is the most energy-consuming device, and its efficiency can be improved by implementing advanced model-based controller design strategies and fault detection and isolation (FDI) methodologies.

Consequently, the development of high-fidelity dynamic centrifugal chiller models has become a priority task of our research. A brief literature review in this field reveals that a significant amount of work has been done on transient and steady state modeling for centrifugal chillers [1]-[4], [7-11].

On the other hand, several dynamic models for vapor compression cycle also have been extensively studied, such as in [6] where is modeled a reciprocating compressor with polytrophic efficiency, that includes also a compressor housing and refrigerant mixing with the oil in the sump. The heat exchanger is modeled

[*]Corresponding Author: Nicolae Tudoroiu, John Abbott College, 2127 Lakeshore Road, Sainte-Anne-de-Bellevue, QC. ntudoroiu@gmail.com

based on moving boundary method that assumes an average state for each phase region according to well mixed modeling assumption. The expansion valve is modeled based on isenthalpic process, and in addition the dynamics of the sensing bulb is also considered. The simulation models are tested with shut-down and start-up and different operations. An interesting approach we find in [9] where a centrifugal compressor model is developed based on control volume methodology that designs the impeller and diffuser separately as control volumes; the impeller is modeled in detail based on the Euler equations, and the diffuser is modeled as ideal without any losses. The mass flow rate and the exit state enthalpy specific for low pressure refrigerant R234a are predicted by two empirical relations. A substantial improvement of this model we find in [10], where a liquid chiller lumped - parameter model of the heat exchangers is developed. Input step changes in the condenser-side water flow rates are studied in order to simulate the system performance under disturbances. Furthermore, in [8] is developed a mechanistic, single stage and two-stage centrifugal chiller models where the centrifugal compressor is modeled based on the Euler turbo-machinery theory, the energy balance and the impeller velocity equations. The coefficient of performance (COP) of the chiller is simulated by considering the compressor polytropic efficiency, hydrodynamic, mechanical and electrical losses. On the other hand, the condenser and evaporator are modeled based on lumped parameter approach and the heat transfer is calculated based on effectiveness model. The chiller model is validated with a water-to-water chiller test facility. The simulations reveal that the chiller dynamic model predicts higher start-up condenser and evaporator pressures than the measurements, more precisely the modeled dynamics seems to appear faster than the experiment results for convergence to the steady-state. Also, the evaporator-side model performs better than the condenser-side in terms of predicting the transient responses. In an another study [2] a dynamic centrifugal liquid chiller model with flooded-type shell-and-tube heat exchangers is developed, where the compressor model is based on a constant speed and its capacity control is achieved by variable inlet guide vane (IGV). An interesting regression model was applied in [8] to determine the maximum capacity condition, i.e., with wide-open IGV, and the actual mass flow rate was then computed based on a linear relationship of the maximum mass flow rate and the IGV position.

During the transient periods, in a numerical simulation environment was found the optimal values of the initial and the overall system charges. Moreover in [8] is mentioned that for simulations purpose, due to complex nature of the refrigeration cycle, it seems that all the existing simulation models developed for centrifugal chillers include some simplifications for some components. For the centrifugal compressor, the dynamic performance was often approximated based on the compressor characteristic map rather than parameterized dynamic models as is mentioned in [1, 4, 6]. Such approximation is sufficient for simulating the steady-state performance of the chiller, whereas may lead to difficulty in simulating transient processes such as startup, load change and shutdown. In [11] is developed a dynamic model for the centrifugal compressor where the model of the heat exchangers is based on lumped parameters approach.

The objective of our research is to develop a comprehensive dynamic model for the centrifugal chiller suitable for control analysis and design. In particular we aim to achieve quality simulation performance in MATLAB and SIMULINK modeling, and simulations environment for all components of centrifugal chiller. The significance of the proposed research includes the following:

- The dynamic characteristics of the centrifugal chiller system was modeled based on detailed mass, momentum and energy balance equations
- The shell-and-tube heat exchangers were modeled based on lumped parameters approach.
- The developed control strategies based on the centrifugal chiller model cover different operating conditions in order to be robust to several changes of the parameters control structure or modeling errors, and to achieve the best transient and steady-state performance.

In this paper we are focused to get accurate models for centrifugal chiller and based on these models to develop a few PI closed-loop control strategies to find the most suitable control strategy for these kinds of applications that performs better in terms of accuracy, robustness, convergence speed, overshoot, and disturbance rejection. The performance comparison of the proposed control strategies is very useful to choose the most suitable control strategy that performs better in terms of meeting the control design requirements and process control constraints, rejection of the effect of possible disturbances that act on the controlled process, as well as a tracking error accuracy.

2. The Centrifugal Chiller Nonlinear Model

2.1. Modeling and description of components

In Figure 1 is shown a schematic of the water-cooled centrifugal chiller system which consists of four major components: a centrifugal compressor, an expansion valve, an evaporator, and a condenser [8]. In general, the overall chilled-water system consists of one circuiting refrigerant loop and two water loops. The first water loop is circuited between the condenser and the cooling tower, and the second water loop is circuited between the evaporator and the air handling unit (AHU). In our research, a detailed nonlinear model of a centrifugal chiller system is developed. The modeling methodology and the model equations are described in Annex-1 appended at the end of this paper. The nonlinear model consists of four major sub-component models, namely, a flooded evaporator, a flooded water-cooled condenser, a centrifugal compressor and an electronic expansion valve. The simulation of transient and steady-state behavior of the water-cooled centrifugal chiller is performed in a well-known MATLAB / SIMULINK environment. As described in Annex-1, the dynamic model of the overall centrifugal chiller system is of high complexity in terms of state dimension and nonlinearity. The nonlinear model equations were then reformulated into a state space model by a set of 40 first order differential equations (ODE), as a mathematical form to represent the physical chiller system.

The significant aspects of the nonlinear model are presented briefly in this section.

First, we focus our attention to centrifugal compressor that is the much faster responding device in the chiller plant. Similar as

is shown in [8] the centrifugal compressor capacity balance is achieved with variable rotor speed control. In our research, the capacity control is manipulated by adjusting the IGV position and input torque, as shown in Figure 2. The model takes into account also, the effects of incidence and fluid friction losses, mentioned in [8]. Figure 2 schematically depicts the centrifugal model with boundary conditions and capacity control. It is important to note that the working medium in the centrifugal chillers is refrigerant that works under multi-phase conditions, which makes it more complex and difficult for the underlying modeling framework and consistent numerical initialization [8].

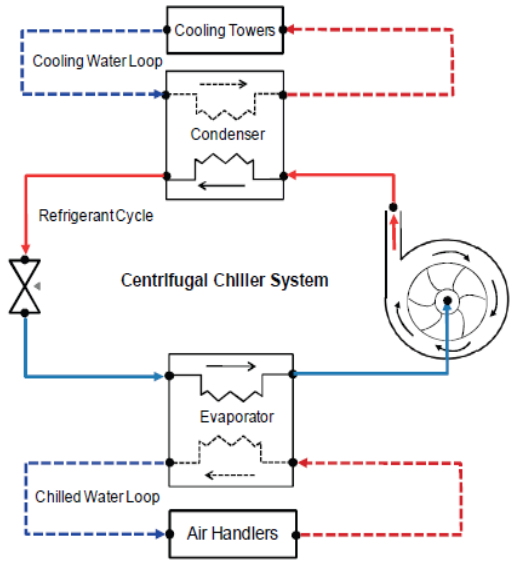


Figure 1. Schematic of water-cooled centrifugal chiller system (see [8])

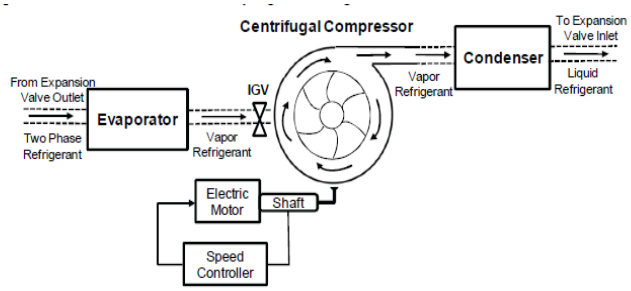


Figure 2. The schematic of the centrifugal compression system in chiller (see [8])

The evaporator and condenser modeling requires quality heat exchanger models. The flooded evaporator was divided into a two-phase section and a superheat section as detailed in Annex-1. Likewise, the flooded condenser was divided into two superheat sections, a two-phase section and a sub-cool section.

The electronic expansion valve (EXV) was modeled to regulate the flow rate of refrigerant in the system and in this study the EXV was used to maintain the liquid level of the refrigerant in the condenser. The modeling details are described in Annex-1.

The dynamic model for centrifugal chiller system presented in Annex-1 is reformulated in state-space representation, by a linear set of ordinary differential equations (ODE), as a natural representation of a physical system, and easily to be solved using a suitable MATLAB solver.

3. MATLAB Open-Loop Simulation results

Simulations were performed with the nonlinear centrifugal chiller model to investigate its open-loop dynamic behavior and steady-state performance using R134a as refrigerant, in a MATLAB/SIMULINK simulation environment. In open-loop the centrifugal chiller is a multi-input, multi-output (MIMO) control system that has two inputs, namely the compressor motor drive speed (Ucom), and the expansion valve opening (U_EXV), two outputs corresponding to chilled water temperature (Tchwsp), and the liquid level in condenser (Level)). Using a consistent set of initial conditions, simulation runs were conducted under constant design load conditions. The open-loop simulation results are shown in Figure 3. The temperature and pressure responses are smooth and reach their respective design operating conditions as the system reaches steady state. The responses in Figure 3 depict the condensing pressure Pc in 3(a), the evaporation pressure Pev in 3(b), chilled water temperature (Tchw,sp) in 3(c), water temperature in the condenser 3(d), the heat exchange rate in 3(e) and the coefficient of performance in 3(f). From these results it can be noted that the system responses reach steady state in about 30 minutes when subjected to constant load. At steady state the coefficient of performance of the system is 5.5 and the heat rejection rate is about 20% greater than the cooling capacity as a result of heat generated by the compressor motor.

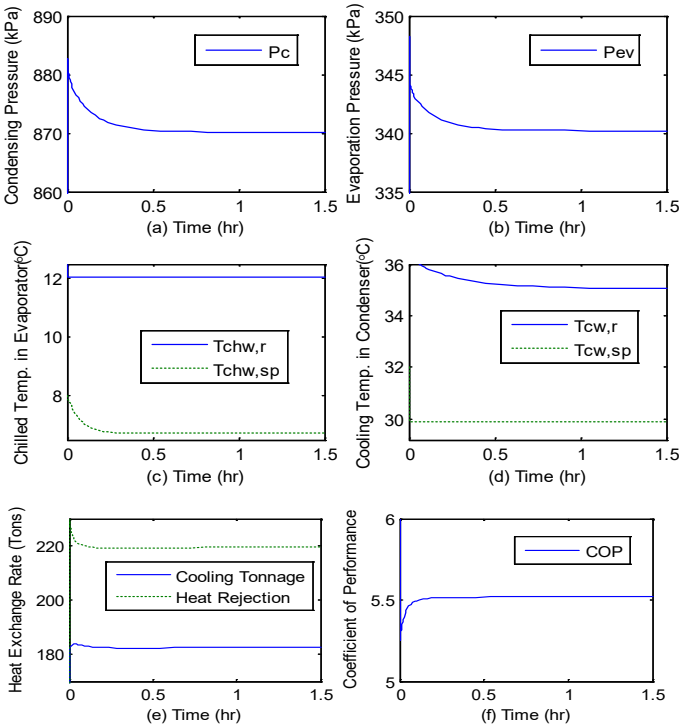


Figure 3. Open-loop responses of the chiller system

Legend: (a) Compression pressure (Pc); (b) Evaporation pressure (Pev); (c)Chilled water temperature Tchw, sp; (d) Water temperature in condenser Tcw; (e) Cooling tonnage and heat rejection; (f) Coefficient of performance (COP);

The SIMULINK model of centrifugal chiller system in open-loop and its details are shown in Figures 4 and 5. These figures with higher resolution are also depicted in Annex-2.

The MATLAB Function1 in the SIMULINK model from Figure 5 has the following code lines:

```
function [U1, U2, Y1, Y2] = fcn(x)
Tchw_sp = x(18);
Level = x(35);
Y1 = Tchw_sp;
Y2 = Level;
U1 = x(33);
U2 = x(37);
end
```

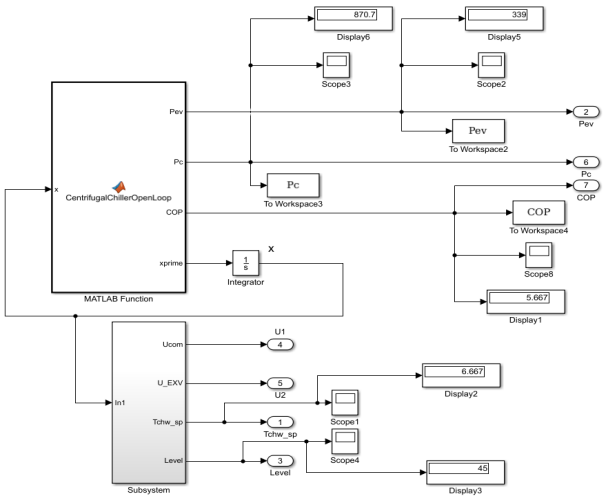


Figure 4: The SIMULINK model of centrifugal chiller in open-loop

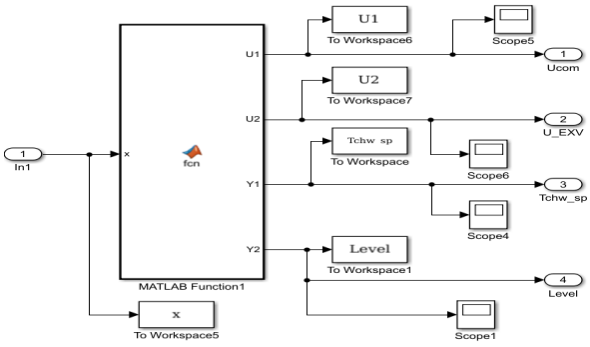


Figure 5. The SIMULINK model of the bottom Subsystem Block

The Centrifugal Chiller open-loop model is very useful to build the PI control strategy as described in the next section for the overall system seen as centralized system, since in "real-life" there exist some interferences between the loops. Furthermore, through open-loop simulations we generate sets of input-output data required to build the ARMAX models for Centrifugal chiller. This is the reason to start with a good open-loop model capable to capture entire dynamics of the overall chiller system under various operating conditions. This give us more flexibility for control design.

4. Closed – Loop Control Strategies Design

In this section we develop and implement in a MATLAB/SIMULINK environment three closed-loop control strategies, namely an overall Proportional-Integral (PI) controller based on the chiller system model in open-loop developed and implemented in MATLAB simulation environment in previous section 3. By extensive simulations we found that the interference between the both loops, temperature and liquid Refrigerant level, respectively, is very week, and thus we can simplify the overall chiller system model by considering that the loops are decoupled.

This is an essential modeling aspect that is investigated in this research paper, by building for each loop separately a PI controller, as is developed in subsection 4.1. The SIMULINK model of PI controllers for each separate loop is shown in Annex-2.

The open-loop model is useful also to generate the set of input-output data measurements necessary to build the linear ARMAX SISO models. The input-output data set measurements file was loaded in a new MATLAB code file that estimates the ARMAX SISO models.

However, from accuracy considerations, in the future research work we will investigate also the extension of ARMAX multi-input single output (MISO) models for building predictive control strategies. In subsection 4.2 the ARMAX SISO decoupled models will be used to build one linear PI controller for each loop separately. The third control strategy is an improved PI control version that is conceived within the context of non-minimum state space control system. The non-minimal state space representation in contrast to minimal state space realizations "seems to be the natural description of a discrete-time transfer function, since its dimension is dictated by the complete structure of the model", as is mentioned in [12-16]. More precisely, we talk about on improved version of Proportional - integral control, so called Proportional-integral-plus (PIP) control for systems with time delay [12-16], that is developed in subsection 4.3. Based on ARMAX models is proved also the superiority of new PPI controller compared with the first two PI versions in terms of efficiency, implementation simplicity, convergence speed and robustness to the load changes disturbance and to the changes in the noise level of measurement sensors. It is worth to mention again the particular structure for this kind of applications of the MIMO centrifugal chiller control system to be split into two single input-single output (SISO) independent closed - loops control subsystems (i.e. the Evaporator chilled water temperature loop to control the Tchwsp, and the Condenser liquid Refrigerant level loop to control the Refrigerant liquid Level. The preset values of the first closed-loop is Tchw_set = 6.67 [°C], and for the second closed-loop is Level set = 45(%) of the maximum level of the liquid Refrigerant inside the Condenser. Also, a PIP closed-loop control is developed in detail in subsection 4.3. Concluding, at the end of section 4 we can decide on a potential controller choice amongst these three control strategies to find the one that performs better in terms of the accuracy, robustness to the changes in the load magnitude, considered as a disturbance, and to the level of measurement noises injected additionally in each forward loop to Tchwsp, and to liquid Level respectively.

4.1. Closed-loop PI Control Strategy of MIMO
4.2. Centrifugal Chiller System

In this subsection we design for each open-loop a PI control strategy given by [12-16]:

- closed-loop PI control strategy of chilled water temperature Tchwsp of Evaporator:

$$U_{com}(t) = k_{p_1}\varepsilon_1(t) + k_{i_1}\int_0^t \varepsilon_1(\tau)d\tau \quad (1)$$

$$\varepsilon_1(t) = y_{1sp} - y_{1mes}(t) = Tchw_{set} - Tchw_{sp}(t) \quad (2)$$

where $\varepsilon_1(t)$ is the error between the set point value of the closed-loop output temperature $Tchw_{set}(t)$ value and its measured value $Tchw_{sp}(t)$; k_{p_1} , k_{i_1} are the tuning proportional, integral coefficients of the first PI controller.

- closed-loop PI control strategy of liquid level inside the condenser:

$$U_{EXV}(t) = k_{p_2}\varepsilon_2(t) + k_{i_2}\int_0^t \varepsilon_2(\tau)d\tau + k_{d_2}\frac{d\varepsilon_2}{dt} \quad (3)$$

$$\varepsilon_2(t) = y_{2sp} - y_{2mes}(t) = Level_{set} - Level(t) \quad (4)$$

where $\varepsilon_2(t)$ is the error between the set point value of the closed-loop output liquid level $Level_{set}(t)$ value and its measured value $Level(t)$; k_{p_1} , k_{i_1} are the tuning proportional, integral coefficients of the second PI controller.

An extensive number of simulations for adjusting the coefficients of both PI controllers in order to perform well in terms of overshoot, settling time and steady-state error led to the following values:

- For closed-loop Evaporator subsystem:

$k_{p_1} = 0.1; k_{i_1} = 0.0001; Tchwsp_set = 6.67[°C]$,

i.e. the Temperature set-point value.

- For closed-loop Condenser subsystem

$k_{p_2} = 0.05; k_{i_2} = 0.0001; Level_set = 45[\%]$, i.e. the Liquid level setting point value.

The closed-loop simulation results related to PI controller performance for Evaporator subsystem are shown in Figure 6 for Condenser subsystem, and in Figure 7 they are related to the overall performance in terms of coefficient of performance COP of entire MIMO control system. In Figure 6 are shown the chilled water Evaporator temperatures (Tchwsp, Tchwsp_set) as in 6(a), the evaporation pressure Pev in 6(b), and the Compressor motor speed actuator efforts in 6(c).

In Figure 7 is showing the PI controller closed-loop performance for Condenser subsystem related to the Refrigerant liquid level, such in 7(a), condensing pressure Pc as is revealed in 7(b), and expansion valve actuator effort presented in 7(c).

The overall performance for entire MIMO control system in the particular case of nonlinear plant dynamics in terms of coefficient of performance is shown in Figure 8.

Concluding, in this section we can see that the closed-loop PI controller of Condenser subsystem perform better compared to the closed-loop PI controller of Evaporator subsystem in terms of accuracy, fast transient, and disturbance rejection, as is shown in Figure 9 for the Evaporator chilled water temperature, and in Figure 10 for Refrigerant liquid level in the Condenser.

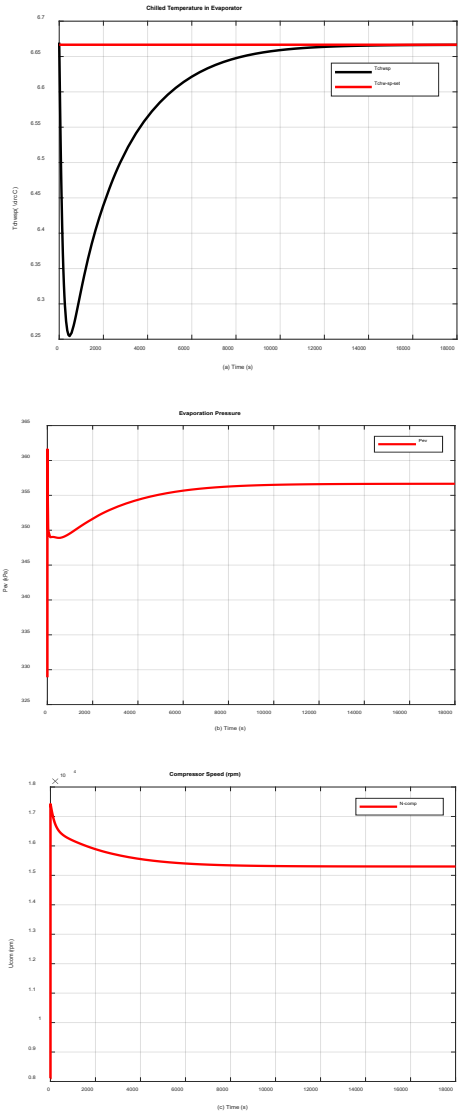


Figure 6. PI control strategy for closed-loop simulations for chiller plant model- Evaporator Subsystem

Legend: (a) Temperature water supply pressure Tchwsp; (b) Evaporation pressure Pev; (c) Compressor motor drive actuator speed

As disturbance, is considered a sharp variation of the load, i.e. the temperature of chilled water Tchw_rr that increases from 48[°C] to 52[°C]), injected inside the MIMO control system by addition in the forward path of the chilled water temperature of Evaporator subsystem before the measured chilled water temperature output, after 9000 seconds (2 hours and 30 minutes), persisting further 2 hours and 30 minutes. This time assures for MIMO control system enough time to reach a new steady-state.

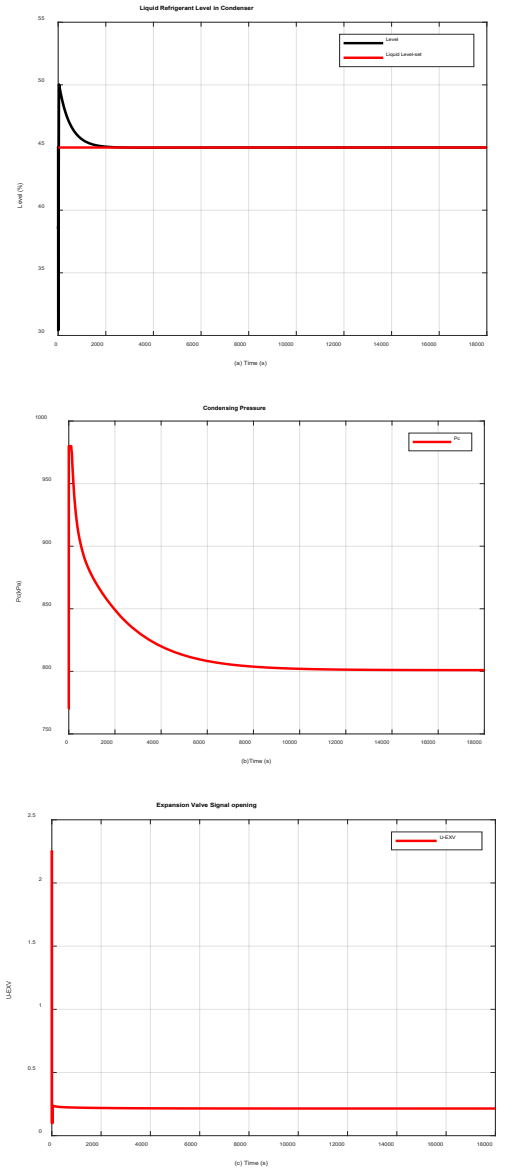


Figure 7. PI control strategy for closed-loop simulations for chiller plant model – Condenser Subsystem

Legend: (a) Refrigerant liquid level; (b) Condensing pressure Pc; (c) Expansion valve actuator opening

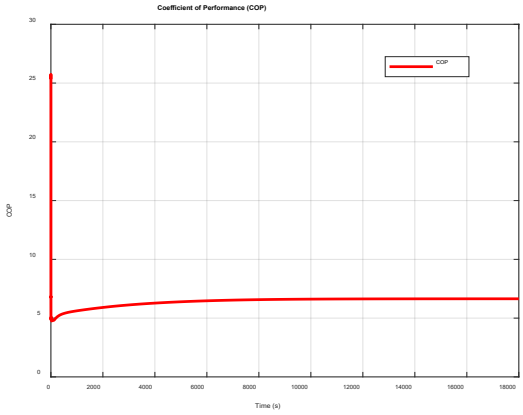


Figure 8 Overall coefficient of performance COP

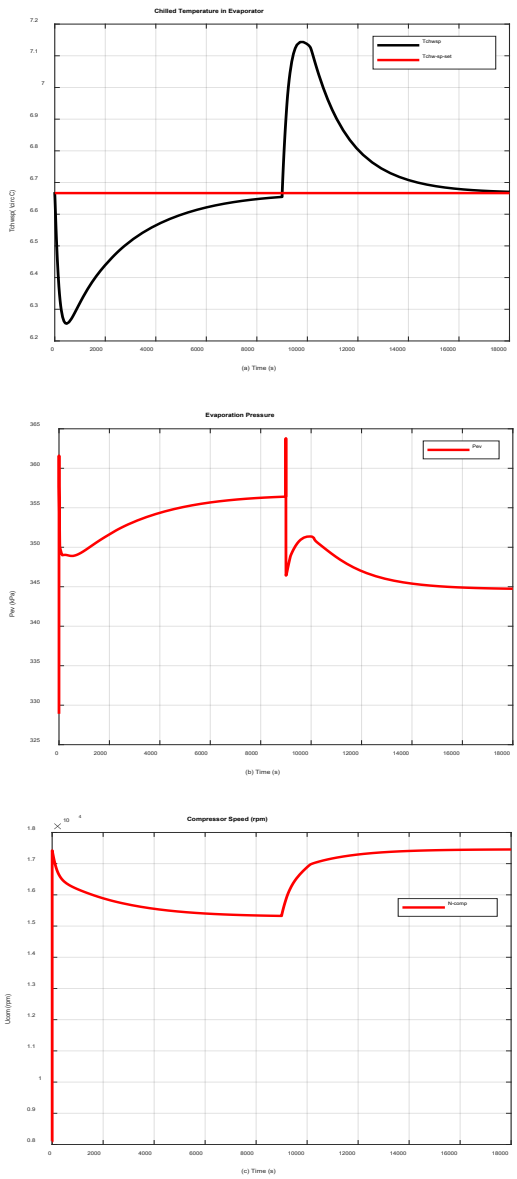


Figure 9. PI control strategy for closed-loop simulations for chiller plant model - Evaporator subsystem, disturbance rejection

Legend: (a) Temperature water supply pressure Tchwsp; (b) Evaporation pressure Pev; (c) Compressor motor drive actuator speed

The both PI controllers perform very well and prove a good robustness to the changes in the control system load, with a significant amount effort of both actuators, especially by the compressor motor drive. The SIMULINK models of both PI control strategies are attached for clearness in the Annex-2. The simulation results reveal also that the PI controller of the Refrigerant liquid level in Condenser subsystem compared to PI controller of the Evaporator subsystem performs faster, during a very short transient time to reject the disturbance impact, therefore is more robust. Despite the changes in the load is worth to mention that the coefficient of performance for entire MIMO control system, as is shown in Figure 11, changes smoothly and regains in short time the optimal value reached in steady-state before the disturbance injection.

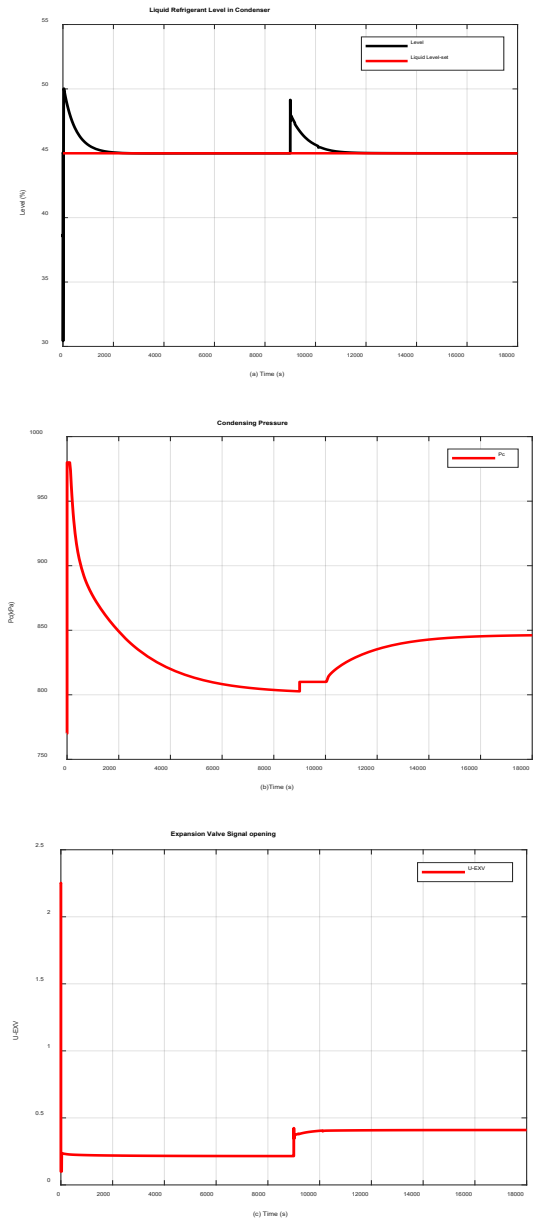


Figure 10. PI control strategy for closed-loop simulations for chiller plant model – Condenser subsystem, disturbance rejection
Legend: (a) Refrigerant liquid level; (b) Condensing pressure Pc; (c) Expansion valve actuator opening U_EXV

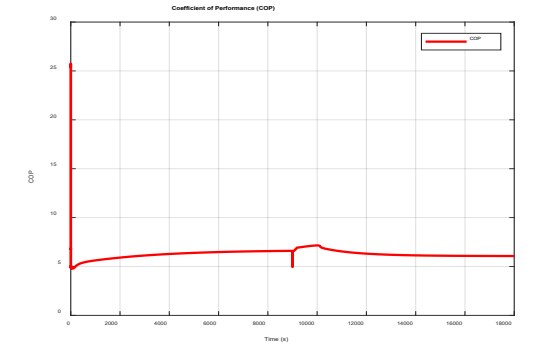


Figure 11. Robustness of the overall coefficient of performance COP to a sharp disturbance variation in the load

Additionally, a normally distributed Gaussian random signal is injected in the output chilled water temperature loop and Refrigerant liquid level loop in order to simulate the impact of the measurement noise level. For a variance in measurement noise level of 0.0001 in chilled water temperature, and 0.1 in the Refrigerant liquid level, their impact on the both PI controllers performance are shown in Figure 12 and Figure 13. The simulation results shown in Figure 13 reveal a good behavior of the closed-loop PI controller of the Refrigerant liquid level compared to closed-loop PI controller of chilled water temperature.

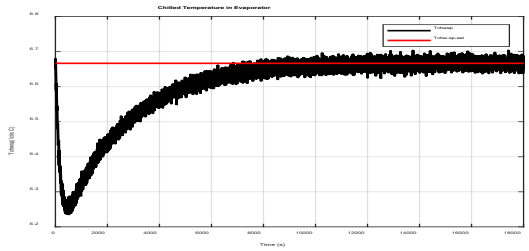


Figure 12. Tchwsp - Measurement noise impact on the MIMO control system response (noise level variance 0.001)

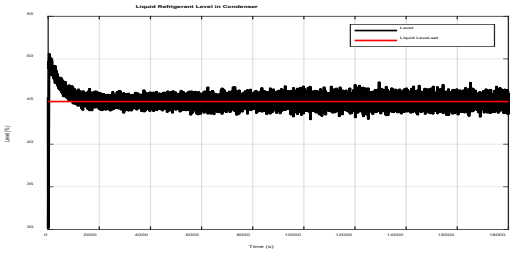


Figure 13. Refrigerant liquid level - Measurement noise on the MIMO control system response (noise level variance 0.1)

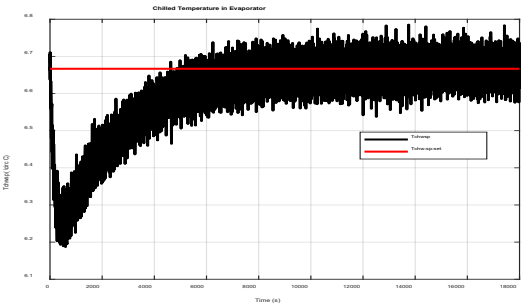


Figure 14 Tchwsp - Measurement noise impact on the MIMO control system response (noise level variance increased to 0.001)

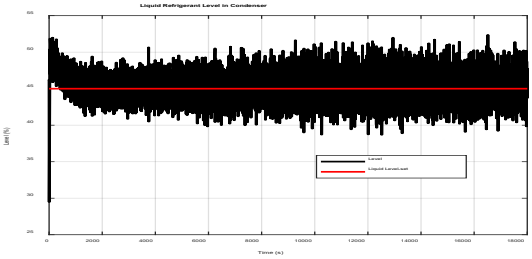


Figure 15. Refrigerant liquid level - Measurement noise impact on the MIMO control system response (noise level variance increased to 1)

The impact due to ten times rise in the measurement noise levels becomes significant, as is shown in Figure 14, and Figure 15. The simulation results reveal a fragile robustness of the both PI controllers to changes in noise variance level.

The noise variance in the measurements can be attenuated by using two Moving Average filters coded in MATLAB, with a windows length of 10, and 100 samples respectively. The filtered responses of MIMO control system are shown in Figure 16 and Figure 17.

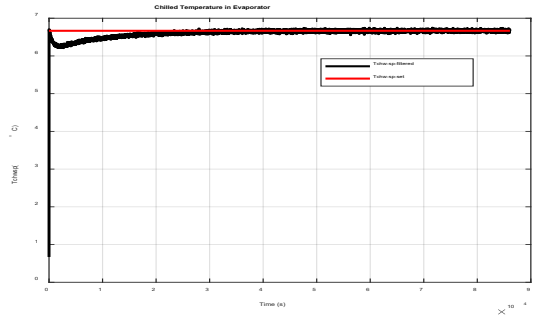


Figure 16. Tchwsp – Moving Average filtered measurement noise impact on the MIMO control system response (noise level variance increased to 0.001, windows length = 10)

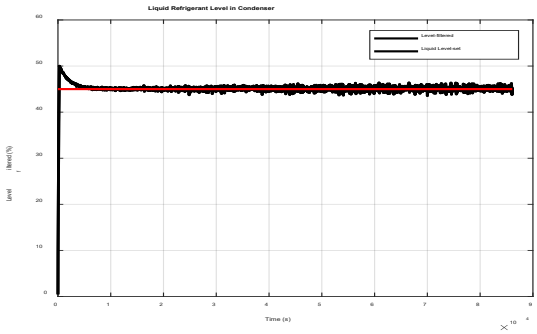


Figure 17. Refrigerant liquid level - Moving Average filtered measurement noise impact on the MIMO control system response (noise level variance increased to 1, window length =100 samples)

4.3. *Closed-loop PI Control Strategies using Linear ARMAX SISO Models for Centrifugal Chiller System*

In this section we develop two linear Auto Regressive Moving Average (ARMAX) SISO models for the both subsystems (Evaporator and Condenser) of the centrifugal chiller system (CCS). These two linear ARMA models are useful in this section to build and implement two PI closed-loop control strategies, and furthermore an interesting approach of a new control design approach, namely a PIP control strategy developed in next subsection 4.3. To estimate the models' parameters are used two particular MTLAB functions from Identification Toolbox, more precisely armax and arx. First one, armax MATLAB function estimates the parameters of ARMAX MIMO or SISO polynomial discrete-time models using time-domain data [17]. The data set of input-output chiller system measurements are generated by open-loop simulations performed on the SIMULINK model of chiller system shown in Figure 4, Figure 5.

The second one, *arx* MATLAB function estimates the parameters of two polynomials discrete-time models known as Autoregressive with an exogenous input (ARX) and a more simple polynomial Autoregressive without exogenous input (AR) that estimates the parameters of the scalar time series, by means of the well-known least squares method the most used in control systems identification and parameters estimation. The both MATLAB functions use a prediction-error method and specified polynomial orders. For each of these models pure transport delays of the signals flow in the feedback path from the measurement sensors to the controllers for each input/output pair are also specified in their ARAMAX and ARX structures [17]. In addition, in each ARMAX and ARX model structures noise channels are incorporated, hence the most general polynomial form of these discrete-time models is given by [18 - 21]:

$$
\begin{aligned}
&A(q)y(t) = B(q)u(t-n_k) + C(q)e(t)\\
&A(q) = 1 + a_1 q^{-1} + a_2 q^{-2} + \ldots + a_{n_a} q^{-n_a} = 1 + \sum_{i=1}^{n_a} a_i q^{-i}\\
&\quad = 1 + q^{-1}\sum_{i=1}^{n_a} a_i q^{-i+1} = 1 + q^{-1}A^*(q^{-1}),\\
&B(q) = b_1 + b_2 q^{-1} + b_2 q^{-2} + \ldots + b_{n_b} q^{-n_b+1} = \sum_{i=1}^{n_b-1} b_i q^{-i}\\
&\quad = q^{-1}\sum_{i=1}^{n_b-1} b_i q^{-i+1} = q^{-1}B^*(q^{-1}),\\
&C(q) = 1 + c_1 q^{-1} + c_2 q^{-2} + \ldots + c_{n_c} q^{-n_c}\\
&\quad = 1 + \sum_{i=1}^{n_c} c_i q^{-i} = 1 + q^{-1}\sum_{i=1}^{n_c} c_i q^{-i+1} = 1 + q^{-1}C^*(q^{-1})
\end{aligned}
\tag{5}
$$

where

- $y(t)$ is the output at discrete-time t, n_a, n_b, n_c represent the degrees of the polynomials $A(q), B(q)$, and $C(q)$ respectively,
- $u(t)$ is the control system input at the discrete instant t,
- n_k is an integer number of sampling periods, the so-called *dead time* and most used in the control systems (pure transport delay of the signal flow between the measurement sensors and controllers in the feedback path,
- a_i, b_i, c_i represent the coefficients (parameters) of the polynomials $A(q), B(q)$, and $C(q)$ of ARMAX models respectively,
- q @ forward shift operator, i.e. $q(y(t)) = y(t+1)$, and q^{-1} @ backward shift operator, i.e. $q^{-1}(y(t)) = y(t-1)$,
- t @ $\frac{t}{T_s}$, denotes the normalized sampling time, only symbolic that replaces the most used typical notation $t = kT_s, k \in \mathbf{Z}^+$ used to designate the description of dynamical discrete-time system, T_s representing the sampling period ,

- $e(t)$ denotes the white noise disturbance value at the discrete instant t

With these notations the general discrete-time representation of a dynamic system as ARMAX model can be put in the following form, as is given in [20]:

$$y(t+1) = -A^*(q^{-1})y(t) + q^{-n_k}B^*(q^{-1})u(t) + (1+q^{-1}C^*(q^{-1}))e(t) \quad (6)$$

If the data are as time series with no input channels and with only one output channel, then armax calculates an ARMA model for the time series, given by [1, 20]:

$$y(t+1) = -A^*(q^{-1})y(t) + (1+q^{-1}C^*(q^{-1}))e(t) \quad (7)$$

Furthermore, if in the noise source channel there is an integrator ARMAX model becomes an ARIMAX model, described by [19]:

$$y(t+1) = -A^*(q^{-1})y(t) + q^{-n_k}B^*(q^{-1})u(t) + \frac{(1+q^{-1}C^*(q^{-1}))}{1-q^{-1}}e(t) \quad (8)$$

If there are no inputs, the ARIMAX structure becomes an ARIMA model of the following form:

$$y(t+1) = -A^*(q^{-1})y(t) + \frac{(1+q^{-1}C^*(q^{-1}))}{1-q^{-1}}e(t) \quad (9)$$

It is worth also to mention that the discrete-time representation in (6) can be put in a so called "regression form" useful to make the link with the least square estimation method of the ARMAX model coefficients and parameters based on the input-output data set measurements, as is mentioned in [19]:

$$y(t+1) = \theta^T\varphi(t) \quad (10)$$

where the vectors θ, φ represent the vector of parameters:

$$\theta = [a_1, a_2, ..., a_{n_a}, b_1, b_2, ..., b_{n_b}]^T \quad (11)$$

and the so called the "*regressor vector*" of measured values of the current outputs and the inputs and theirs the past values, respectively:

$$\varphi(t) = [-y(t), -y(t-1), ..., -y(t-n_a+1), u(t-n_k), u(t-n_k-1), ..., u(t-n_k-n_b+1)] \quad (12)$$

By means of some symbolic algebraic manipulations performed in the first equation (5) we get an interesting result that makes the link with the well-known discrete time "*transfer operator*" for open or closed loop control systems:

$$y(t) = q^{-n_k}\frac{B(q^{-1})}{A(q^{-1})}u(t) + \frac{C(q^{-1})}{A(q^{-1})}e(t) = G_u(q^{-1})u(t) + G_e(q^{-1})e(t) \quad (13)$$

$$G_u(q^{-1}) = q^{-n_k}\frac{B(q^{-1})}{A(q^{-1})}, G_e(q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})} \quad (14)$$

where $G_u(q^{-1})$ and $G_e(q^{-1})$ represents the discrete time transfer operators for the channels input-output control system in open or closed-loop, $u(t) \to y(t)$ and "*white*" or "*colored*" noise channel – output control system, $e(t) \to y(t)$ respectively.

Moreover, if we replace formally in the equations (13) and (14) the shift (backward or forward) operator q by a complex variable $z \in C, z = e^{st}, s = \sigma + j\omega, \sigma = \mathrm{Re}(s), \omega = \mathrm{Im}(s)$ representing the real and imaginary parts of the complex variable s respectively, the most used by Laplace transform to compute the transfer functions of the control systems in the complex domain s, a new representation in z complex domain is found, as follows:

$$Y(z) = z^{-n_k}\frac{B(z^{-1})}{A(z^{-1})}U(z) + \frac{C(z^{-1})}{A(z^{-1})}E(z) = G_u(z^{-1})U(z) + G_e(z^{-1})E(z) \quad (15)$$

$$G_u(z^{-1}) = z^{-n_k}\frac{B(z^{-1})}{A(z^{-1})}, G_e(z^{-1}) = \frac{C(z^{-1})}{A(z^{-1})} \quad (16)$$

This description is typically used for the control systems represented in discrete-time, where $U(z) @Z\{u(t)\}, Y(z) @Z\{y(t)\}$, and $E(z) = Z\{e(t)\}$ are the z-images of the input $u(t)$, white noise $e(t)$, and the output $y(t)$, obtained by applying the "transform Z" operator to all three discrete - time variables $u(t)$, $e(t)$, and $y(t)$ respectively [19]. The transfer operators $G_u(z^{-1})$, and $G_e(z^{-1})$ defined in equation (16) are so-called z-transforms functions of the discrete-time control system.

In our case study, the discrete-time representations of the both SISO ARMAX models assigned to the centrifugal chiller control system are obtained in MATLAB R2017b by some manipulations of typical MATLAB functions in MATLAB code, given by [18-21]:

A. Chilled water temperature closed-loop SISO ARMAX discrete-time model:

$$y_1(t) = q^{-2}\frac{B(q^{-1})}{A(q^{-1})}u_1(t) + \frac{C(q^{-1})}{A(q^{-1})}e_1(t) = G_{u_1}(q^{-1})u(t) + G_{e_1}(q^{-1})e_1(t) \quad (17)$$

$$G_{u_1}(q^{-1}) = q^{-2}\frac{B(q^{-1})}{A(q^{-1})}, G_{e_1}(q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})} \tag{18}$$

and the polynomials $A(q^{-1}), B(q^{-1})$ are given by:

$$A(q^{-1}) = 1 - 2.909q^{-1} + 2.822q^{-2} - 0.9131q^{-3} \tag{19}$$

$$B(q^{-1}) = 0.01843 - 0.005826q^{-1} - 0.001592q^{-2} \tag{20}$$

where the polynomials $A(q^{-1}), B(q^{-1}),$ and $C(q^{-1})$ have the orders 3, 3, and 6 respectively. The pure transport delay is $n_k = 2$ samples, so the polynomials coefficients are given by:

$$\begin{aligned} &a_{1T} = -2.909, a_{2T} = 2.822, a_{3T} = -0.9131 \\ &b_{1T} = 0.01843, b_{2T} = -0.005826, b_{3T} = -0.001592 \end{aligned} \tag{21}$$

B. Refrigerant liquid level closed-loop SISO ARMAX discrete-time model:

$$\begin{aligned} y_2(t) &= q^{-2}\frac{B(q^{-1})}{A(q^{-1})}u_2(t) + \frac{C(q^{-1})}{A(q^{-1})}e_2(t) \\ &= G_{u_2}(q^{-1})u_2(t) + G_{e_2}(q^{-1})e_2(t) \end{aligned} \tag{22}$$

$$G_{u_2}(q^{-1}) = q^{-2}\frac{B(q^{-1})}{A(q^{-1})}, G_{e_2}(q^{-1}) = \frac{C(q^{-1})}{A(q^{-1})} \tag{23}$$

$$A(q^{-1}) = 1 - 0.9938q^{-1} \tag{24}$$

$$B(q^{-1}) = 1.129q^{-1} + 0.152q^{-2} \tag{25}$$

$$C(q^{-1}) = 1 + 0.7744q^{-1} \tag{26}$$

$$\begin{aligned} &a_{1L} = -0.9938 \\ &b_{1L} = 1.129, b_{2L} = 0.152 \\ &c_{1L} = 0.7744 \end{aligned}$$

In Figure 18 we gather a valuable information about the validation of the both ARMAX models for entire range of the input-output measurements data set. To validate these two models in MATLAB can be used the typical MATLAB function called *compare* found in Identification MATLAB Toolbox. The input-output data set measurement of length 3006 samples are generated in open-loop simulations based on SIMULINK model of chiller system developed in section 3. The first segment of 1800 samples from input-output data set is used in prediction phase, and the second segment containing the samples between 1800 and 3006 is used in validation phase. The simulation results shown in Figure 18 reveal a very good estimation accuracy of the both ARMAX SISO models.

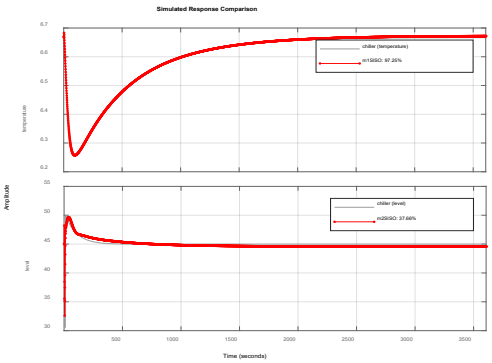


Figure 18. Estimated ARMA SISO model validation for the booth closed-loops, chilled water temperature and Refrigerant liquid level

The closed-loop PI control laws for the both SISO ARMAX models are similar to those developed in the equations (1) - (4), but their design is more suitable in discrete-time to match the both SISO ARMAX models. The both PI closed loops control strategies are built in SIMULINK and the simulations results are performed partially in SIMULINK and finalized in MATLAB, as is shown in Figure 19, and Figure 20. In Figure 19 are shown the simulations results for chilled water temperature in Evaporator closed-loop control subsystem, and in Figure 20 can be seen the simulation results of closed-loop PI controller performance for Refrigerant liquid level in Condenser subsystem. The both PI control loops are completely independent, without any interference between them. They are based on the decoupling loops assumption, proved in open-loop MATLAB simulation environment. The simulation results reveal good accuracy and fast transient, especially for Refrigerant liquid level PI control shown in Figure 20. For the first approximate 500 samples the chilled water temperature in Evaporator reaches low values compared to measured temperature from the input-output measurements data set required for the estimation, prediction and validation of both ARMAX models. Fortunately, after this starting transient period the controlled chilled water temperature reaches the output measurements data set values with high accuracy. This is an interesting modelling key issue generated by the linearization of the plant dynamics, leading for a short period of time during the transient to a considerable degradation in PI controller performance. In addition the constraints on the operating range of the actuators could be also a big issue for PI controller design. In contrast, the simulation results from Figure 20 reveal a very good accuracy, and also a fast transient.

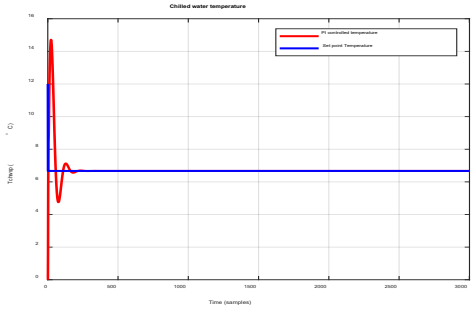


Figure 19. PI controlled chilling water Evaporator temperature closed-loop ARMA SISO model compared to the set point temperature value

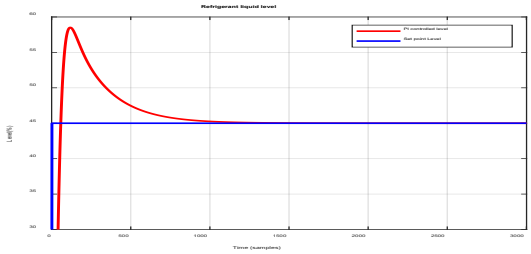


Figure 20. PI controlled liquid level closed-loop ARMA SISO model compared

Set point level value

The Simulink models for the both PI controllers based on ARMAX SISO models are shown in Figure 21 and Figure 22, and also for more clearness they are attached in Annex-2. The rejection disturbance performance of both PI controllers is shown in Figure 23 and Figure 24.

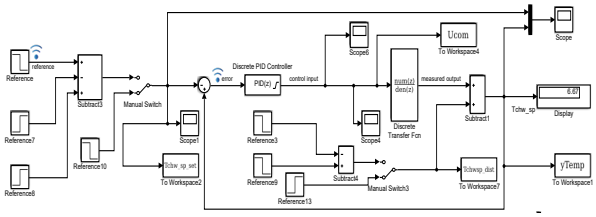


Figure 21. The SIMULINK model of chilled water temperature PI closed-loop Evaporator control subsystem

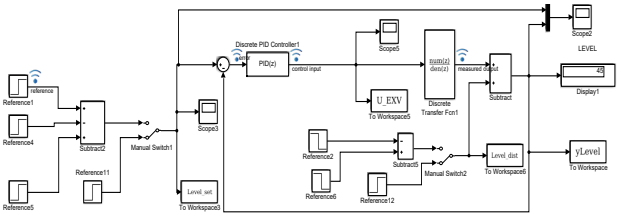


Figure 22. The SIMULINK model of the Refrigerant liquid level PI closed-loop Condenser control subsystem

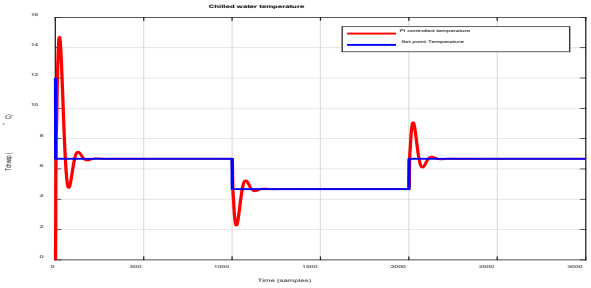


Figure 23. Robustness of PI controlled chilling water Evaporator temperature closed-loop ARMA SISO model to the changes in set point temperature values

The simulations results in Figure 23 and Figure 24 show a great load disturbance rejection performance for both PI controllers. In fact, these results prove also the robustness of the both PI controllers for changes in the chiller system load.

Concluding, compared to the both PI control strategies developed in subsection 4.1, the proposed PI controllers in this subsection perform successful when using linear ARMAX SISO models for both control subsystems Evaporator and Condenser.

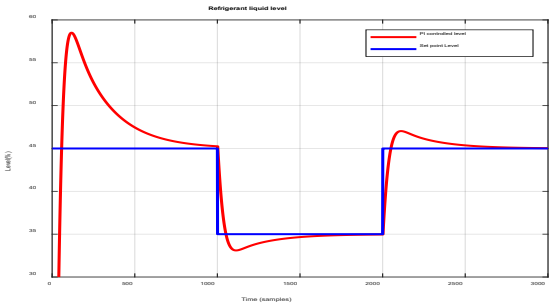


Figure 24 Robustness of PI controlled refrigerant liquid level in Condenser closed-loop ARMA SISO model to the changes in set point level values

4.4. *PPI Closed-Loop Controller Design using Linear ARMAX Models for Centrifugal Chiller Control System with Time Delay*

The novelty of this subsection is a new design approach of the standard PI control strategy as a predictive discrete-time PI-plus (PIP) controller within the context of non-minimum state space (NMSS) control system, as is formulated in [13]. Furthermore, the non-minimal state space (NMSS) representation in contrast to minimal state space realizations "*seems to be the natural description of a discrete-time transfer function, since its dimension is dictated by the complete structure of the model*" as is stated in [13]. The minimal state space descriptions account only for the order of the denominator of the transfer function, and also the state variables assigned to each description not always have physical meaning, usually representing combinations of input and output signals [13]. Therefore, the resulting control algorithm in the new PIP control design approach within the NMSS context "*can be interpreted as a logical extension of the conventional PI controller, facilitating its straightforward implementation using a standard hardware-software arrangement*", as is stated also in [13]. Further, the controller design methodology is the same as for an equivalent discrete-time Smith predictor (SP) controller for time delay systems, under certain non-restrictive pole assignment conditions, as is mentioned in [13].

The proposed PPI controller design follows the same methodology described in [13], encouraged by its great results obtained when applied in a MIMO ALSTOM nonlinear gasifier control plant [15]. Compared to the well-known Smith Predictor controller, in the new design approach the predictive PI-plus controller (PIPC) has more design flexibility and robustness [13].

Furthermore, the tuning parameters optimal selection, such as the weighting matrices required in the optimal performance criterion formulation can achieve multiple objectives. In addition, this new approach can be easy extended to MIMO PIPCs based on ARMAX models, such in [15]. In our case study the centrifugal chiller control system is represented in discrete-time by two SISO loops ARMAX models, given by the equations (17) - (21), and also in z-domain described by the equations (15) –(16). To simplify further the description in z-domain we neglect the white noise term, and the equations (15) and16) can be written in the following compact form, similar to those introduced in [13]:

$$y(k)=\frac{(b_{n_k}z^{-1}+b_{n_{k+1}}z^{-2}+b_{n_{k+2}}z^{-3}+...+b_{n_{k+m-1}}z^{-m})z^{-n_k+1}}{1+a_1z^{-1}+a_2z^{-2}+...+a_nz^{-n}}u(k) \quad (27)$$
$$=\frac{B(z^{-1})z^{-n_k+1}}{A(z^{-1})}=H_y^u(z)$$

In (27) $H_y^u(z)$ represents the transfer function in z-domain of the output y of centrifugal chiller control SISO loop system with respect to corresponding control input u. Following the design control methodology described in [14], the model of the control system defined by the discrete transfer function given in (27) can be represented by the following NMSS equations:

$$x(k)=Fx(k-1)+gu(k-1)+dy_{sp}(k) \quad (28)$$

$$y(k)=hx(k) \quad (29)$$

where the $n+m+n_k-1$ dimensional NMSS state vector $x(k)$ is defined as follows:

$$x(k)=[y(k)\ y(k-1)...y(k-n-1)\ u(k-1)...u(k-m-n_k+2)\ z(k)]^T \quad (30)$$

The new variable $z(k)$ introduced in (30) represents the discrete-time integral of error between the set point input (reference) $y_{sp}(k)$ and the control system output $y(k)$:

$$z(k)=z(k-1)+(y_{sp}(k)-y(k)) \quad (31)$$

The matrix F, and the vectors g,d,h are calculated based on the methodology described in [13-17]. Due to the editing constraints we give below only the matrix F and the vectors g,d,h for a delay $n_k=2$ since matches the SISO temperature loop, and if is the case, for $n_k\geq 2$ you can see the general description used in [13] for SISO systems, that can be also extended for MIMO systems. Similar, in [15] is investigated the case $n_k=1$ for a MIMO system that can be tailored on the Refrigerant liquid level SISO loop in our case study. For SISO systems with a delay $n_k=2$ that matches also the temperature SISO loop, the matrix F and the vectors g,d,h can be written as following [13]:

$$F=\begin{bmatrix} -a_1 & -a_2 & ... & -a_{n-1} & -a_n & b_{n_k} & ... & b_{n_k+m-2} & b_{n_k+m-1} & 0 \\ 1 & 0 & ... & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & ... & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & ... & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ . & . & ... & . & . & . & ... & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ a_1 & a_2 & ... & a_{n-1} & a_n & -b_{n_k} & ... & -b_{n_k+m-2} & -b_{n_k+m-1} & 1 \end{bmatrix} \quad (32)$$

$$g=[0\ 0\ \ldots 0\ 1\ 0\ 0.....0\ 0\ 0]^T$$
$$=[0\ 0\ \ldots 0\ g_i=1\ 0\ 0.....0\ 0\ 0]^T$$
$$d=[0\ 0\ \ldots 0\ 0\ 0\ 0\ 0...0\ 0\ 1]^T \quad (33)$$
$$h=[1\ 0\ \ldots 0\ 1\ 0\ 0.....0\ 0\ 0]$$

The element $g_i=1$ of the vector g occupies the same position as the component $u(k-1)$ occupies in the state vector $x(k)$ defined by (30). The components of the vectors d,h are identified straightforward due to their simple structures. A MIMO control system characterized by a transport delay $n_k=1$ is described in [15] by the following expressions:

$$G=\begin{bmatrix} B_1 & Z_p & Z_p & ... & Z_p & I_p & Z_p & ... & Z_p & Z_p & -B_1 \end{bmatrix}^T$$
$$D=\begin{bmatrix} Z_p & Z_p & Z_p & ... & Z_p & Z_p & Z_p & ... & Z_p & Z_p & I_p \end{bmatrix}^T \quad (34)$$
$$H=\begin{bmatrix} I_p & Z_p & Z_p & ... & Z_p & Z_p & Z_p & ... & Z_p & Z_p & Z_p \end{bmatrix}^T$$

$$I_p=\begin{bmatrix} 1 & 0 & . & 0 \\ 0 & 1 & . & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{bmatrix}_{p\times p}, \quad Z_p=O_{p\times p}$$

$$F=\begin{bmatrix} -A_1 & -A_2 & ... & -A_{n-1} & -A_n & B_2 & B_3 & ... & B_{m-1} & B_m & 0 \\ I_p & Z_p & ... & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p \\ Z_p & I_p & ... & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p \\ . & . & . & . & . & . & . & . & . & . & . \\ Z_p & Z_p & ... & I_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p \\ Z_p & Z_p & ... & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p & Z_p \\ Z_p & Z_p & ... & Z_p & Z_p & I_p & Z_p & Z_p & Z_p & Z_p & Z_p \\ Z_p & Z_p & ... & Z_p & Z_p & Z_p & I_p & Z_p & Z_p & Z_p & Z_p \\ . & . & ... & . & . & . & . & . & . & . & . \\ I_p & Z_p & ... & Z_p & Z_p & Z_p & Z_p & Z_p & I_p & Z_p & Z_p \\ A_1 & A_2 & ... & A_{n-1} & A_n & B_2 & B_3 & ... & B_{m-1} & B_m & I_p \end{bmatrix} \quad (35)$$

The control law $u(k)$ related to the NMSS realization (28-30) takes the following state variable feedback (SVF) form [14-16]:

$$u(k)=-k_F x(k) \quad (36)$$

$$k_F=[f_0\ \ f_1\ \ f_2...f_{n-1}\ g_1\ g_2.....g_{m+n_k-2}\ \ -k_I]^T \quad (37)$$

where the last component k_I of k_F is the integrator gain of the first block located in the forward path of the PIP control diagram shown in Figure 25.

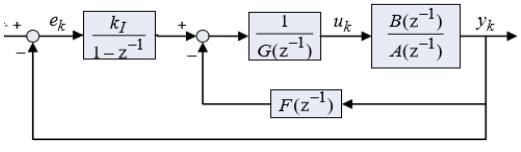


Figure 25. PIP control block diagram (see [16])

It is worth to mention that the linear SVF control law (33) is easy to be implemented in practice, due to the facility storage of the variables in the MATLAB workspace. Furthermore, the inherent SVF formulation allows to investigate any SVF design methodology, such as:

- Poles assignment
- Linear quadratic (LQ) optimization
- Linear Quadratic Gaussian optimization
- H_∞ optimization
- Linear exponential of quadratic Gaussian (LEQG) optimization

The poles assignment PIP controller design methodology in many cases is much simpler and intuitive, since the algebraic results are more perceptibly. In this paper we are focused only on the second approach, linear quadratic (LQ) optimization, but for the interested readers we try to provide only the steps of the poles assignment PIP controller design methodology, as follows:

Step1. Define a characteristic polynomial for the desired behavior of closed-loop PIP control system, i.e. choose the roots of the desired characteristic polynomial inside of unit circle $|z|<1$ to assure the stability of the PIP control system in closed-loop, and to reach the desired performance (poles assignment).

The desired characteristic polynomial that meet the requirements from first step can be written as follows:

$$L(z^{-1}) = 1 + l_1 z^{-1} + l_2 z^{-2} + ... + l_{n+m+n_k-1} z^{-n-m-n_k+1} \quad (38)$$

Step2. Compute the discrete time transfer function in closed-loop for the PIP control system that has the block diagram presented in Figure 24, as is given in [13]:

$$H_y^{y_{sp}}(z) = \frac{k_I B(z^{-1}) z^{-n_k+1}}{(1-z^{-1})[G(z^{-1})A(z^{-1}) + F(z^{-1})B(z^{-1})z^{-n_k+1}] + k_I B(z^{-1}) z^{-n_k+1}}$$

$$H_y^{y_{sp}}(z) = \frac{y(k)}{y_{sp}(k)} = \frac{B_{cl}(z^{-1})}{A_{cl}(z^{-1})}, B_{cl}(z^{-1}) = k_I B(z^{-1}) z^{-n_k+1}, \quad (39)$$

$$A_{cl}(z^{-1}) = (1-z^{-1})[G(z^{-1})A(z^{-1}) + F(z^{-1})B(z^{-1})z^{-n_k+1}] + k_I B(z^{-1}) z^{-n_k+1}$$

Step 3. Identify the polynomials coefficients $f_i|_{i=\overline{0,n-1}}$, $g_i|_{i=\overline{1,m+n_k-2}}$ of $F(z^{-1})$, and $G(z^{-1})$ respectively, as well as the integrator gain k_I:

$$F(z^{-1}) = f_0 + f_1 z^{-1} + ... + f_{n-1} z^{-n+1} \quad (40)$$

$$G(z^{-1}) = g_0 + g_1 z^{-1} + ... + g_{m+n_k-2} z^{-m-n_k+2} \quad (41)$$

Identification of the polynomials coefficients is based on the following poles assignment equation:

$$A_{cl}(z^{-1}) = L(z^{-1}) \quad (42)$$

assuming that the closed-loop desired poles are located inside the unit circle to assure the stability, a good transient, as well as a good set point tracking.

The LQ optimization design approach is applied to the both closed-loops SISO ARMAX models (19-20), (25-26), considered as a starting point in order to design two PIP control laws.

Typically, the optimization criterion of LQ optimization design is defined in the same manner as for any quadratic optimization form written for a SISO control system:

$$J = \sum_{k=0}^{\infty} (x(k)^T Q x(k) + r u(k)^2) \quad (43)$$

where Q is a diagonal weighting matrix of the form:

$$Q = diag([q_1 \quad q_2 ..... \quad q_i \quad ..... \quad q_{n-1} \quad q_n]), q_{i|i=\overline{1,n}} > 0 \quad (44)$$

and $r > 0$ is a positive scalar weight on the scalar input u.

The resulting SVF gains are then obtained recursively from the steady state solution of the Algebraic Riccati Equation (ARE), derived from the standard LQ cost function (14) as follows [16]:

$$k_F(k) = [g^T P^{(k+1)} g + r]^{-1} g^T P^{(k+1)} F \quad (45)$$

$$P^{(k)} = F^T P^{(k+1)} [F - g k_F] + Q \quad (46)$$

where the matrix P is a symmetrical positive definite matrix with its initial value $P^{(0)} = Q$, and k_F is the control gain vector for SVF.

A. PIP control law for chilled water temperature closed-loop SISO ARMAX discrete-time Evaporator model described by:

$$A(z^{-1}) = 1 - 2.909 z^{-1} + 2.822 z^{-2} - 0.9131 z^{-3}$$

$$B(z^{-1}) = 0.01843 z^{-1} - 0.01436 z^{-2} - 0.004013 z^{-3}$$

$$A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3},$$

$$a_1 = -2.909, a_2 = 2.822, a_3 = -0.9131$$

$$B(z^{-1}) = b_2 z^{-1} + b_3 z^{-2} + b_4 z^{-3}, b_2 = 0.01843,$$

$$b_3 = -0.01436, b_4 = -0.004013$$

$$n = 3, m = 3, n_k = 2$$

$$F(z^{-1}) = f_0 + f_1 z^{-1} + f_2 z^{-2}$$

$$G(z^{-1}) = 1 + g_1 z^{-1} + g_2 z^{-2} + g_3 z^{-3}$$

The state associated to the PPI controller is given by:

$$x(k) = [y(k) \quad y(k-1) \quad y(k-2) \quad u(k-1) \quad u(k-2) \quad u(k-3) \quad z(k)]^T$$

and the PIP controller parameters computed according to (32) and (33) are given by:

$$F = \begin{bmatrix} -a_{1T} & -a_{2T} & -a_{3T} & b_{2T} & b_{3T} & b_{4T} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ a_{1T} & a_{2T} & a_{3T} & -b_{2T} & -b_{3T} & -b_{4T} & 1 \end{bmatrix}$$

$$g = [0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0]^T$$
$$k_{sp} = [0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1]^T$$
$$h = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

The tuning parameters are set to:

$$Q = \begin{bmatrix} 150 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \text{and } r = 1.$$

The simulation results are shown in Figure 26 that reveal a very good tracking performance, fast transient, no oscillations, and a very small overshoot, perhaps the best performance obtained until now compared to previous temperature controllers.

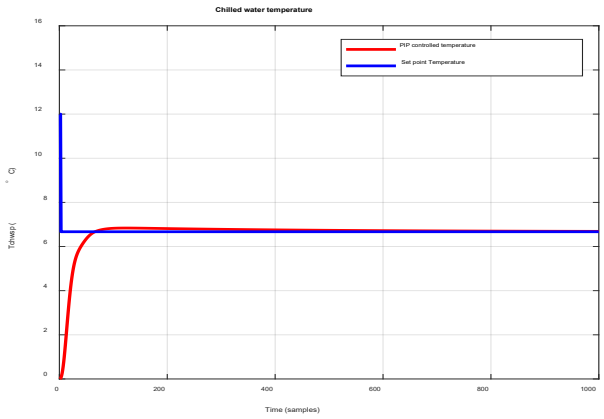


Figure 26. PIP controlled chilling water Evaporator temperature closed-loop ARMA SISO model compared to the set point temperature value

B. PIP control law for closed-loop Refrigerant liquid level SISO ARMAX discrete-time Condenser model described by:

$$A(q^{-1}) = 1 - 0.9938q^{-1}$$
$$B(q^{-1}) = 1.129q^{-1} + 0.152\,q^{-2}$$
$$C(q^{-1}) = 1 + 0.7744q^{-1}$$
$$a_{1L} = -0.9938$$
$$b_{1L} = 1.129, b_{2L} = 0.152$$
$$c_{1L} = 0.7744$$

$$Z_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \text{zeros}(2), I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \text{eye}(2)$$

The polynomials degree is smaller compare to temperature loop description, i.e. $n = 2, m = 2, n_k = 1$, thus the state vector will be of small dimension that simplify very much the PIP controller design, such as:

$$x(k) = [y(k) \quad u(k-1) \quad z(k)]^T$$

and the model parameters are calculated according to (34) replacing $I_p = I_1 = 1, Z_p = Z_1 = 0$ to match to an ARMAX SISO closed-loop control, as follows:

$$F = \begin{bmatrix} -a_{1L} & b_{2L} & 0 \\ 1 & 0 & 0 \\ a_{1L} & -b_{2L} & 1 \end{bmatrix}, g = [b_{1L} \quad 1 \quad -b_{1L}]^T$$
$$\text{ksp} = [0 \quad 0 \quad 1]^T, \text{h} = [1 \quad 0 \quad 0]$$

and the tuning parameters are set to

$$Q = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}, r = 1$$

The Simulink models of PIP controllers are shown in Figure 27 for chilled water temperature control in Evaporator subsystem, and Figure 28 for liquid Refrigerant level control in Condenser subsystem, respectively.

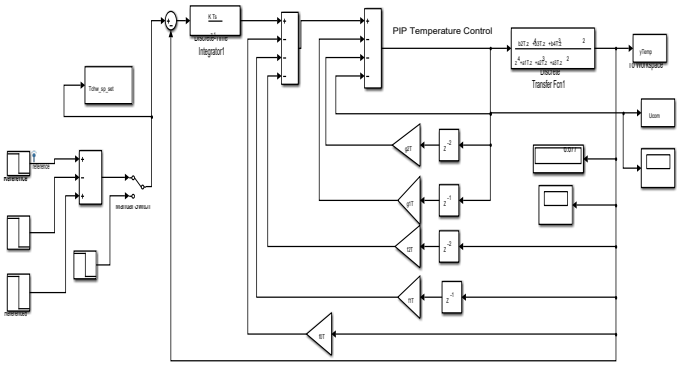


Figure 27. The PIP controller of chilled water temperature in Evaporator Subsystem

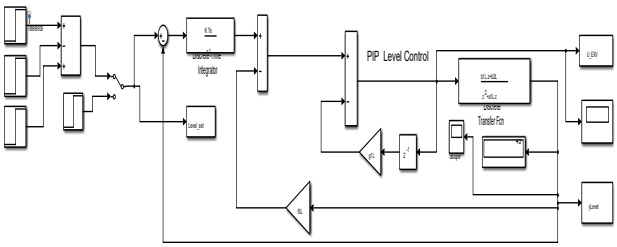


Figure 28. The PIP controller of the liquid Refrigerant level in Condenser Subsystem

The simulation results are shown in Figure 29 that reveal a very fast transient and no oscillations, thus no overshoot and the overall behavior of PIP controller is as an aperiodic element. The same conclusion we can formulate that is the best performance obtained until now compared to previous Refrigerant liquid level controllers. Furthermore, we test the both PIP controllers for robustness to changes in set points, and the results reveal an excellent tracking performance for both, as is shown in Figure 30, and Figure 31 respectively.

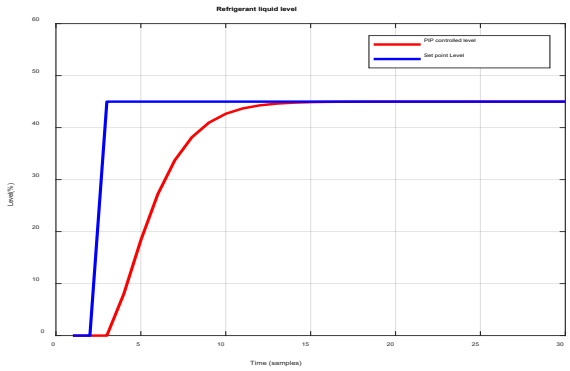


Figure 29. PIP controlled Refrigerant liquid level closed-loop ARMA SISO model compared to the set point level value

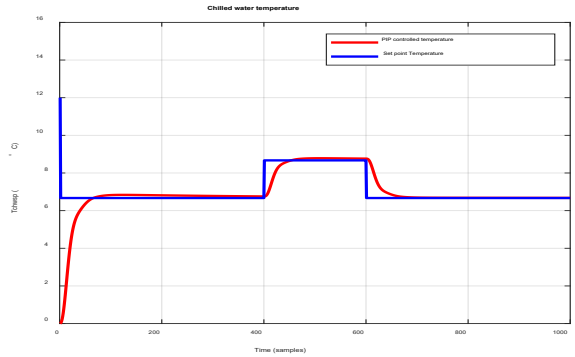


Figure 30.The robustness to changes in input set point for closed-loop PIP controlled chilled water temperature in Evaporator subsystem described by an ARMA SISO model

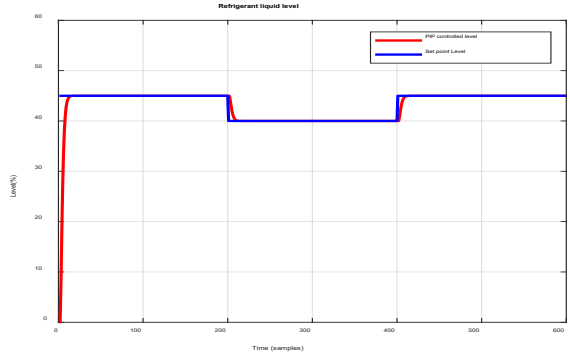


Figure 31. The robustness to changes in input point set for closed-loop PIP controlled Refrigerant liquid level in Condenser subsystem described by an ARMA SISO model

Conflict of Interest

The authors declare no conflict of interest.

Conclusions

In our research work we introduce a new approach controller design, the so called in the literature as PI controller plus [13-16], definitely a new version of PI controller, but much simpler for design, and no more tuning parameters are required compared to standard version of PI controller, i.e. the coefficients k_p, k_I. The third control strategy is an improved PI control version conceived within the context of non-minimum state space control systems with time delay.

The non-minimal state space representation in contrast to minimal state space realizations "seems to be the natural description of a discrete-time transfer function, since its dimension is dictated by the complete structure of the model", as is stated in [13]. Also, overall the new PIP controller developed in subsection 4.3 outperforms the first two PI controller's versions proposed in the subsection 4.1 and 4.2 in terms of tracking performance, robustness, convergence speed and overshoot. Furthermore, the PIP controller can be easily extended to control MIMO systems with delay, as is done in [15], thus remains an open research topic for further investigations in the future work.

References

[1] S. Bendapudi, J. E. Braun, et al., "Dynamic Model of a Centrifugal Chiller System - Model Development, Numerical Study, and Validation", *ASHRAE Trans.*, vol. 111, pp.132-148, 2005.

[2] A. Beyene, H. Guven, et al., (1994)," Conventional Chiller Performances Simulation and Field Data", *International Journal of Energy Research*, vol.18, pp.391-399, 1994.

[3] J. E. Braun, J. W. Mitchell, et al., "Models for Variable-Speed Centrifugal Chillers", *ASHRAE Trans.* , New York, NY, USA, 1987.

[4] M. W. Browne, P. K. Bansal, "Steady-State Model of Centrifugal Liquid Chillers", *International Journal of Refrigeration*, vol. 21, no.5, pp. 343-358, 1998.

[5] B.E. Deepa Th. Mannath, "Study and Simulation of the Predictive Proportional Integral Controller in comparison with Proportional Integral Derivative controllers for a two-zone heater system", Master Thesis, Texas Tech University, 2002.

[6] M. Dhar, W. Soedel, "Transient Analysis of a Vapor Compression Refrigeration System", in *The XV International Congress of Refrigeration*, Venice, 1979.

[7] J.M. Gordon, K. C. Ng, H.T. Chua, "Centrifugal chillers: thermodynamic modeling and a diagnostic case study", *International Journal of Refrigeration*, vol.18, no. 4, pp.253-257, 1995.

[8] Li Pengfei, Li Yaoyu, J. E. Seem, "Modelica Based Dynamic Modeling of Water-Cooled Centrifugal Chillers", *International Refrigeration and Air Conditioning Conference*, Purdue University, pp.1-8, 2010. *http://docs.lib.purdue.edu/iracc/1091*.

[9] P. Popovic, H. N. Shapiro, "Modeling Study of a Centrifugal Compressor", *ASHRAE Trans.,* Toronto, 1998.

[10] M. C. Svensson, "Non-Steady-State Modeling of a Water-to-Water Heat Pump Unit", in *Proceedings of 20th International Congress of Refrigeration*, Sydney, 1999.

[11] H. Wang, S. Wang, "A Mechanistic Model of a Centrifugal Chiller to Study HVAC Dynamics", in *Building Services Engineering Research and Technology*, vol. 21, no.2, pp. 73-83, 2000.

[12] A. A, Nada, E. M. Shaban, "The Development of Proportional-Integral-Plus Control Using Field Programmable Gate Array Technology Applied to Mechatronics System", American Journal of Research Communication, vol. 2, no.4, pp.14-27, 2014.

[13] C J Taylor, A Chotai and P C Young, "Proportional-integral-plus (PIP) control of time delay systems", in *Proceedings of the Institute Mechanical of Engineers*, vol. 212, no.1, pp.37-47, 1998.

[14] P. C. Young, A. P. McCabe, A. Chotai "State-dependent parameter nonlinear systems: Identification, Estimation and Control", in *IFAC, 15th Triennial World Congress*, Barcelona, Spain, 2002.

[15] C. J. Taylor and E.M. Shaban, "Multivariable Proportional-Integral-Plus (PIP) control of the ALSTOM nonlinear gasifier simulation", in *IEE Proceedings Control Theory and Applications*, vol. 153, no.3, pp.277–285, 2006.

[16] E. M. Shaban, A. A. Nada, "Proportional Integral Derivative versus Proportional Integral plus Control Applied to Mobile Robotic System", *Journal of American Science*, vol.9, no.12, pp.583-591, 2013, doi:10.7537/marsjas091213.76.

[17] M. Zaheeruddin, N. Tudoroiu, "Neuro - PID tracking control of a discharge air temperature system", Elsevier, *Energy Conversion and Management*, vol. 45, pp.2405–2415, 2004.

[18] MATLAB R2017b Documentation, *https://www.mathworks.com/help/ident/ref/armax.html.* Accessed at February 3rd, 2018.

[19] I.D. Landau, R. Lozano, M. M'SAAD, A. Karimi, *Adaptive Control – Chapter 2: Discrete-Time System Models for Contro*l, Springer Link, pp. 35-53, 2011, DOI: 10.1007/978-0-85729-664-1_2.
[20] L. Ang, "Comparison between Model Predictive Control and PID Control for water-level maintenance in a two-tank system", Master Thesis, University of Pittsburgh, 2010.
[21] Carnegie Mellon Lab, University of Michigan, "Control Tutorials for MATLAB", http://ctms.engin.umich.edu.
[22] M. Ning, "Neural Network Based Optimal Control of HVAC&R Systems", PHD Thesis, Department of Building, Civil and Environmental Engineering. Montreal, Quebec: Concordia University, 2008.
[23] ASHRAE Handbook-HVAC Systems and Equipment, 2008.
[24] S.A. Korpela, Principles of turbomachinery. A John Wiley & Sons, Inc., Publication. 2011.
[25] S.P.W. Wong and S.K. Wang," System Simulation of the performance of a Centrifugal chiller Using a Shell-and-Tube-Type Water-Cooled Condenser and R-11 as Refrigerant", in ASHRAE Trans, pp.445-454, 1989.
[26] F.W. Yu and K. T. Chan, "Improved Condenser Design and Condenser-Fan Operation for Air-Cooled Chillers", Applied Energy vol. 83, pp. 628-648, 2006.

Annex-1. Nonlinear Model for Centrifugal Chiller System

1.1. Development of the centrifugal chiller model

A schematic diagram of a centrifugal chiller system is shown in Figure 1. The main components of the system include: a flooded evaporator; a flooded water-cooled condenser; a centrifugal compressor and an electronic expansion valve. These components and flow direction of the fluids, including the water and the refrigerant loops are depicted in Figure 1.

In order to simplify the model, the following assumptions are made [22-26]:

1) The heat exchange between the chiller and the ambient air is neglected;
2) 15% of the total refrigerant mass flow is used for cooling the hermetic compressor motor;
3) The refrigerant distributes in the shell side of the evaporator and condenser evenly in superheated state, or saturated mixture state of liquid and vapour, or sub-cooled state.
4) The pressure loss due to refrigerant flow in the tubes, discharge line and suction elbow due to friction are neglected, which means that the evaporation pressure is considered as the suction pressure of the compressor, and condensing pressure is regarded as the discharge pressure of the compressor.
5) Hot gas bypass is closed during the operation and simulation.

By applying the fundamental principles of mass, momentum and energy balances, sets of equations for the component models and the overall system model were developed. These are described in the following sections.

1.1.1. Flooded evaporator model

A flooded evaporator is commonly used in centrifugal chiller systems. In the evaporator, the refrigerant R134a flows in the shell side and the chilled water flows inside the tubes in a two-pass circuit. According to the state of the refrigerant, a moving boundary exists, which divides the evaporator into two sections, a two-phase section (TP) where the refrigerant is a mixture of liquid and vapor, and a superheat section (SH) where the refrigerant is superheated. Based on the energy and mass conservation principle, the heat exchange between the chilled water and refrigerant are expressed by the following equations:

$$C_{chw,tp,1}\frac{dT_{chw,1}}{dt} = \dot{m}_{chw,tp,1}\, c_w (T_{chw,r} - T_{chw,1}) - N_{tb,ev,tp,1} U_{ev,tp,} LMTD_{ev,tp,1} \tag{1}$$

$$C_{chw,tp,2}\frac{dT_{chw,sp,1}}{dt} = \dot{m}_{chw,tp,2}\, c_w (T_{chw,1} - T_{chw,sp,1}) - N_{tb,ev,tp,2} U_{ev,tp} LMTD_{ev,tp,2} \tag{2}$$

$$C_{chw,sh}\frac{dT_{chw,sp,2}}{dt} = \dot{m}_{chw,sh}\, c_w (T_{chw,1} - T_{chw,sp,2}) - N_{tb,ev,sh} U_{ev,sh} LMTD_{ev,sh} \tag{3}$$

$$T_{chw,sp} = \frac{N_{tb,ev,tp,2} T_{chw,sp,1} + N_{tb,ev,tp,1} T_{chw,sp,2}}{N_{tb,ev,tp,2} + N_{tb,ev,tp,1}} \tag{4}$$

$$C_{re,ev,sh}\frac{dT_{re,ev,out}}{dt} = N_{tb,ev,sh} U_{ev,sh} LMTD_{ev,sh} - \dot{m}_{re,ev}\, c_{p,re,v} (T_{re,ev,out} - T_{re,tp,ev}) \tag{5}$$

where $\dot{m}_{re,ev}$ is the refrigerant vapour generation rate in the evaporator, kg/s, which is calculated from (6).

$$\dot{m}_{re,ev} = \frac{N_{tb,ev,tp,1} U_{ev,tp,} LMTD_{ev,tp,1} + N_{tb,ev,tp,2} U_{ev,tp,} LMTD_{ev,tp,2}}{h_{gh,Pev}} + x_v (\dot{m}_{re,EXV} + \dot{m}_{re,motor}) \tag{6}$$

In addition, from the mass conservation principle, the following equation was derived for the refrigerant flowing in the shell [22]:

$$\begin{aligned} \dot{m}_{re,EXV} + \dot{m}_{re,motor} - \dot{m}_{re,com} &= \int_0^{Vtp,ev} \frac{\partial \rho_{re}}{\partial t} dV + \int_{Vtp,ev}^{Vev} \frac{\partial \rho_{re}}{\partial t} dV \\ &= \int_0^{Vtp,ev} \frac{\partial (\overline{\gamma_{v,ev}} \rho_{re,satv,Pe} + (1 - \overline{\gamma_{v,ev}}) \rho_{re,satl,Pe})}{\partial t} dV + \int_{Vtp,ev}^{Vev} \frac{\partial \rho_{re,sh,Pe}}{\partial t} dV \end{aligned} \tag{7}$$

According to the Leibniz integral rule,

$$\frac{d}{dz}\left(\int_{x_1(z)}^{x_2(z)} f(x,z)dx\right)=\int_{x_1(z)}^{x_2(z)}\frac{\partial f}{\partial z}dx+f(x_2(z),z)\frac{dx_2}{dz}-f(x_1(z),z)\frac{dx_1}{dz} \tag{8}$$

and by assuming that the density of the refrigerant vapor in the SH section is equal to the saturated vapor density at evaporator pressure, equation(7) can be rewritten as:

$$\begin{aligned}&\dot{m}_{re,EXV}+\dot{m}_{re,motor}-\dot{m}_{re,com}=\\&\left(\frac{d}{dt}\left(\int_0^{Vtp,ev}[\overline{\gamma_{v,ev}}\rho_{re,satv,Pev}+(1-\overline{\gamma_{v,ev}})\rho_{re,satl,Pev}]dV\right)-\rho_{re,satv,Pev}\frac{dV_{tp,ev}}{dt}\right)\\&+\left(\frac{d}{dt}\left(\int_{Vtp,ev}^{Vtp}\rho_{re,v}dV\right)+\rho_{re,satv,Pev}\frac{dV_{tp,ev}}{dt}\right)\\&=(V_{ev}-V_{tp,ev}+\overline{\gamma_{v,ev}}V_{tp,ev})\frac{d\rho_{re,satv,Pev}}{dP_{ev}}\frac{dP_{ev}}{dt}+(1-\overline{\gamma_{v,ev}})V_{tp,ev}\frac{d\rho_{re,satl,Pev}}{dP_{ev}}\frac{dP_{ev}}{dt}\end{aligned} \tag{9}$$

From equation (9), we can get the relationship between the evaporation pressure and the mass of the refrigerant inside the evaporator.

The total number of tubes $N_{tb,ev,tp,1}$, $N_{tb,ev,tp,2}$ and $N_{tb,ev,tp,sh}$ in the two-phase section 1, two-phase section 2 and superheat section respectively are determined by the tubes distribution in the evaporator vessel and the refrigerant mass in the evaporator. By neglecting the refrigerant mass stored in the circulating pipes and compressors, and by assuming that the tubes are distributed evenly in the lower half of evaporator vessel, $N_{tb,ev,tp,1}$, $N_{tb,ev,tp,2}$ and $N_{tb,ev,tp,sh}$ can be determined as follows:

$$V_{net,ev}=0.125\pi D_{ev}^2L_{ev}-0.25\pi N_{tb,ev,total}D_{tb}^2L_{ev} \tag{10}$$

$$V_{tp,ev}=\frac{m_{re,charge}-m_{re,c}}{\overline{\overline{\gamma_{v,ev}}}\rho_{re,satv,Pev}+(1-\overline{\overline{\gamma_{v,ev}}})\rho_{re,satl,Pev}} \tag{11}$$

$$N_{tb,ev}=\frac{V_{re,ev}}{V_{net,ev}}N_{tb,ev,total} \tag{12}$$

if $N_{tb,ev}>N_{tb,ev,total}$, $N_{tb,ev,tp,1}=N_{tb,ev,tp,2}=0.5N_{tb,ev,total}$, $N_{tb,ev,sh}=0$;

else if $0.5N_{tb,ev,total}<N_{tb,ev}<N_{tb,ev,total}$, $N_{tb,ev,tp,1}=0.5N_{tb,ev,total}$, $N_{tb,ev,tp,2}=N_{tb,ev}-N_{tb,ev,tp,1}$, $N_{tb,ev,sh}=N_{tb,ev,total}-N_{tb,ev}$;

else if $N_{tb,ev}<0.5N_{tb,ev,total}$, $N_{tb,ev,tp,1}=N_{tb,ev}$, $N_{tb,ev,tp,2}=0$, $N_{tb,ev,sh}=N_{tb,ev,total}-N_{tb,ev}$.

1.1.2. Flooded condenser model

Based on the mass and energy conservation principle, the heat transfer between the refrigerant and the cooling water inside the condenser was described by the following dynamic equations:

$$C_{cw}\frac{dT_{cw,r}}{dt}=N_{tb,1,c}U_{tb,sh,c}LMTD_{sh,1,c}-\frac{2N_{tb,1,c}}{N_{total,c}}\dot{m}_{cw}c_w(T_{cw,r}-T_{cw,1}) \tag{13}$$

$$C_{cw}\frac{dT_{cw,sh}}{dt}=N_{tb,2,c}U_{tb,sh,c}LMTD_{sh,2,c}-\frac{2N_{tb,2,c}}{N_{total,c}}\dot{m}_{cw}c_w(T_{cw,sh}-T_{cw,sp}) \tag{14}$$

$$C_{cw}\frac{dT_{cw,tp}}{dt}=N_{tb,3,c}U_{tb,tp,c}LMTD_{tp,c}-\frac{2N_{tb,3,c}}{N_{total,c}}\dot{m}_{cw}c_w(T_{cw,tp}-T_{cw,sp}) \tag{15}$$

$$C_{cw}\frac{dT_{cw,sc}}{dt}=N_{tb,4,c}U_{tb,sc,c}LMTD_{sc,c}-\frac{2N_{tb,4,c}}{N_{total,c}}\dot{m}_{cw}c_w(T_{cw,sc}-T_{cw,sp}) \tag{16}$$

$$C_{cw}\frac{dT_{cw,1}}{dt}=N_{tb,2,c}U_{tb,sh,c}LMTD_{sh,2,c}+N_{tb,3,c}U_{tb,tp,c}LMTD_{tp,c}+N_{tb,4,c}U_{tb,sc,c}LMTD_{sc,c}-\frac{2(N_{tb,2,c}+N_{tb,3,c}+N_{tb,4,c})}{N_{total,c}}\dot{m}_{cw}c_w(T_{cw,1}-T_{cw,sp}) \tag{17}$$

$$C_{re,sc}\frac{dT_{re,sc,c}}{dt}=(\dot{m}_{re,EXV}+\dot{m}_{re,motor})c_{p,\mathrm{Re},l,c}(T_{re,tp,c}-T_{re,sc,c})-N_{tb,4,c}U_{tb,sc,c}LMTD_{sc,c} \tag{18}$$

Note that the saturated refrigerant temperature $T_{re,tp,c}$ is determined by the condensing pressure P_c which will be described in the next section. Furthermore, the number of tubes $N_{tb,1,c}$, $N_{tb,2,c}$, $N_{tb,3,c}$ and $N_{tb,4,c}$ in the1st superheat section, the 2nd superheat section, the two-phase section and the sub-cool section, respectively, are determined by the mass of the refrigerant and the tube distribution in the condenser, which will be described later.

1.1.3. Centrifugal compressor model

The percent volumetric flow Θ and the percent head of the refrigerant through the compressor are modeled as in reference [23]:

$$\Omega=\mu(V_p^2/a^2) \tag{19}$$

$$\Theta=\frac{V_p}{a\pi}\frac{\dot{Q}}{U_{com}N_{comd}D^3} \tag{20}$$

where U_{com} is the input signal of the compressor rotational speed, which is used to control the chilled water supply temperature. Friction losses in the pipes and in the evaporator were neglected. In other words, the evaporation pressure is considered as equal to the suction pressure of the compressor by neglecting the pressure loss across the inlet guide vane (IGV). Similarly, the condensing pressure is assumed to be equal to the discharge pressure.

In addition, the polytropic efficiency [24], the polytropic compression work [1] per unit mass of refrigerant, and the motor efficiency [25] were modeled using the approach cited in the above references:

$$\eta_p=(\frac{n-1}{n})(\frac{\gamma}{\gamma-1}) \tag{21}$$

$$W_P=(\frac{P_{disc}}{\rho_{re,disc}}-\frac{P_{suc}}{\rho_{re,suc}})\frac{\ln(P_{disc}/P_{suc})}{\ln[(P_{disc}\rho_{re,suc})/(P_{suc}\rho_{re,disc})]} \tag{22}$$

$$\eta_{motor} = \alpha_1 + \alpha_2(PLR) + \alpha_3(PLR)^2 \tag{23}$$

From the above equations, the total electrical power input to the compressor is computed from (24).

$$P_{elec} = \frac{\dot{m}_{re,com} W_p}{\eta_P \eta_{motor}} \tag{24}$$

1.1.4. Electronic expansion valve (EXV) model

EXV mainly regulates the flow rate of the liquid refrigerant entering the evaporator to maintain the refrigerant mass balance in the flooded evaporator and flooded condenser. In practice, EXV is used to maintain the liquid level of the refrigerant in the condenser. When the liquid level is higher than the set-point, the EXV opening will increase; and *vice versa*. In the absence of EXV physical characteristics and performance data, the mass flow rate was modeled as a function of opening area as described in reference [26].

$$\dot{m}_{re,EXV} = U_{EXV} C_d A_{EXV,\max} \sqrt{2(P_c - P_{ev})\rho_{re,c,out}} \tag{25}$$

By using the mass balance relationship between the total refrigerant mass inside the condenser, including the liquid refrigerant and the refrigerant mixture, and the liquid level L_{evel}, the total number of tubes in each section of the condenser is calculated as follows:

$$m_{re,c} = m_{0,c} + \int_0^t \dot{m}_{re,com} dt - \int_0^t \dot{m}_{re,EXV}\, dt - \int_0^t \dot{m}_{re.motor} dt$$

$$= [\overline{\gamma_{v,c}} \rho_{re,satv,Pc} + (1 - \overline{\gamma_{v.,c}}) \rho_{re,satl,Pc}] \frac{N_{tb,3,c}}{N_{tb,total,c}} V_{net,c} + \rho_{re,satl,Pc} \frac{N_{tb,4,c}}{N_{tb,total,c}} V_{net} \tag{26}$$

$$V_{L,c} = \beta_1 L_{evel}^{\ 3} + \beta_2 L_{evel}^{\ 2} + \beta_3 L_{evel} + \beta_4 \tag{27}$$

$$N_{tb,3,c} + N_{tb,4,c} = N_{tb,total,c} - N_{tb,1,c} - N_{tb,2,c} = \frac{V_{L,c}}{V_{net,c}} N_{tb,total,c} \tag{28}$$

$N_{tb,1,c} = 0.5 N_{tb,total,c}$, while $L_{evel} \le 50\%$

$$N_{tb,3,c} = \frac{\dot{m}_{re,com} h_{re,disc} - N_{tb,1,c} U_{tb,sh,c} LMTD_{sh,1,c} - N_{tb,2,c} U_{tb,sh,c} LMTD_{sh,2,c}}{U_{tb,tp,c} LMTD_{tp,c}} \tag{29}$$

$$N_{tb,2,c} = N_{tb,total,c} - \{(m_{0,c} + \int_0^t \dot{m}_{re,com} dt - \int_0^t \dot{m}_{re,EXV}\, dt - \int_0^t \dot{m}_{re.motor} dt) N_{tb,total,c} / V_{net,c} - N_{tb,3,c} [\overline{\gamma_{v,c}} \rho_{re,satv,Pc} + (1 - \overline{\gamma_{v.,c}}) \rho_{re,satl,Pc}]\} / \rho_{re,satl,Pc} - N_{tb,3,c} - N_{tb,1,c} \tag{30}$$

In addition, by assuming an isenthalpic process across the valve, and by applying the energy and mass balance principles, the enthalpy of the refrigerant entering the evaporator is computed from:

$$h_{re,ev,in} = \frac{\dot{Q}_{motor}}{0.15 \dot{m}_{re,com}} + h_{re,c,out} \tag{31}$$

The model equations were solved using a consistent set of initial conditions. A single compressor chiller system with a cooling capacity of 190-ton was simulated. The open-loop responses of the chiller system at the design operating conditions are presented in Figure 3.

Nomenclature

$A_{EXV,\max}$	maximum opening area of the EXV, m^2
C	thermal capacity, J/℃
C_d	discharge coefficient of the expansion valve
c_w	specific heat of the water, J/(kg*℃)
D	diameter, m
L	length of the evaporator vessel or condenser vessel, m
g	gravitational acceleration, m/s^2
h	rise in pressure head of the refrigerant across the compressor, m
h_{fg}	enthalpy change per unit mass of refrigerant due to phase change, J/kg
h_{re}	enthalpy of the refrigerant, J/kg
L_{evel}	refrigerant liquid level in the condenser
$LMTD$	log-mean temperature difference between the water and the refrigerant, ℃
$\dot{m}$	mass flow rate of the fluid, kg/s
$m_{0,c}$	initial mass of the refrigerant in the condenser, kg
$m_{re,c}$	total mass of the refrigerant in the condenser, kg
$m_{re,charge}$	total mass of the refrigerant charged in the chiller, kg
$\dot{m}_{re,com}$	refrigerant mass flow rate compressed by the compressor, kg/s
$\dot{m}_{re,ev}$	refrigerant vapour generation rate in the evaporator, kg/s
$\dot{m}_{re,EXV}$	refrigerant mass flow rate through the electronic expansion valve, kg/s
$\dot{m}_{re,motor}$	refrigerant mass flow rate used to cool the compressor motor, kg/s
n	polytropic index
$N_{com,\max}$	maximum rotational speed of the compressor, rpm
N_{tb}	number of cropper tubes
P_c	condensing pressure, kPa
P_{elec}	electrical power consumption of the compressor, kW

P_{ev} evaporation pressure, kPa

PLR partial load ratio

ΔP_{re} pressure rise of the refrigerant across the compressor, m

$\dot{Q}_{motor}$ heat generation rate of the compressor, W

t time, s

T temperature, ℃

T_{re} refrigerant temperature, ℃

U overall heat transfer coefficient between the water and the refrigerant per unit length of copper tube, W/℃

U_{IGV} Inlet Guide Vane opening position

U_{EXV} Expansion valve opening position

U_{com} compressor rotational speed signal

v_{disc} specific volume of the refrigerant vapor at the compressor discharge, kg/m^3

V_{ev} total volume of the shell side in the evaporator, m^3

$V_{L,c}$ total volume of the liquid refrigerant and saturated mixture, m^3

V_p tip speed of the compressor impeller, m/s

$\dot{V}_{re}$ volume flow rate of the refrigerant, m^3/s

$V_{tp,ev}$ total refrigerant volume of the two-phase section, m^3

W_P polytropic work input per unit mass of refrigerant, kW/kg

x_v ratio of the vapor refrigerant mass to the total refrigerant mass in the mixture before entering the evaporator

Greek letters

$\overline{\gamma_v}$ average volume fraction of the vapour refrigerant in the TP section

ρ_{re} density of the refrigerant, kg/m^3

β angel of the impeller, rad

γ specific heat ratio of the refrigerant

η_P polytropic efficiency

η_{motor} compressor motor efficiency

Subscripts

1 1st section 2 2nd section

c condenser cw cooling water

chw chilled water ev evaporator

imp impeller
out outlet
Pc condensing pressure
Pev evaporation pressure
r return water
re refrigerant
satv saturated refrigerant vapour
satl saturated refrigerant liquid
sc subcool section
sh superheat section
sp supply water
suc suction of the compressor
total the total number of the tubes
tp two phase section

Annex-2. SIMULINK models of open-loop and PI and PIP closed-loop control of chiller system

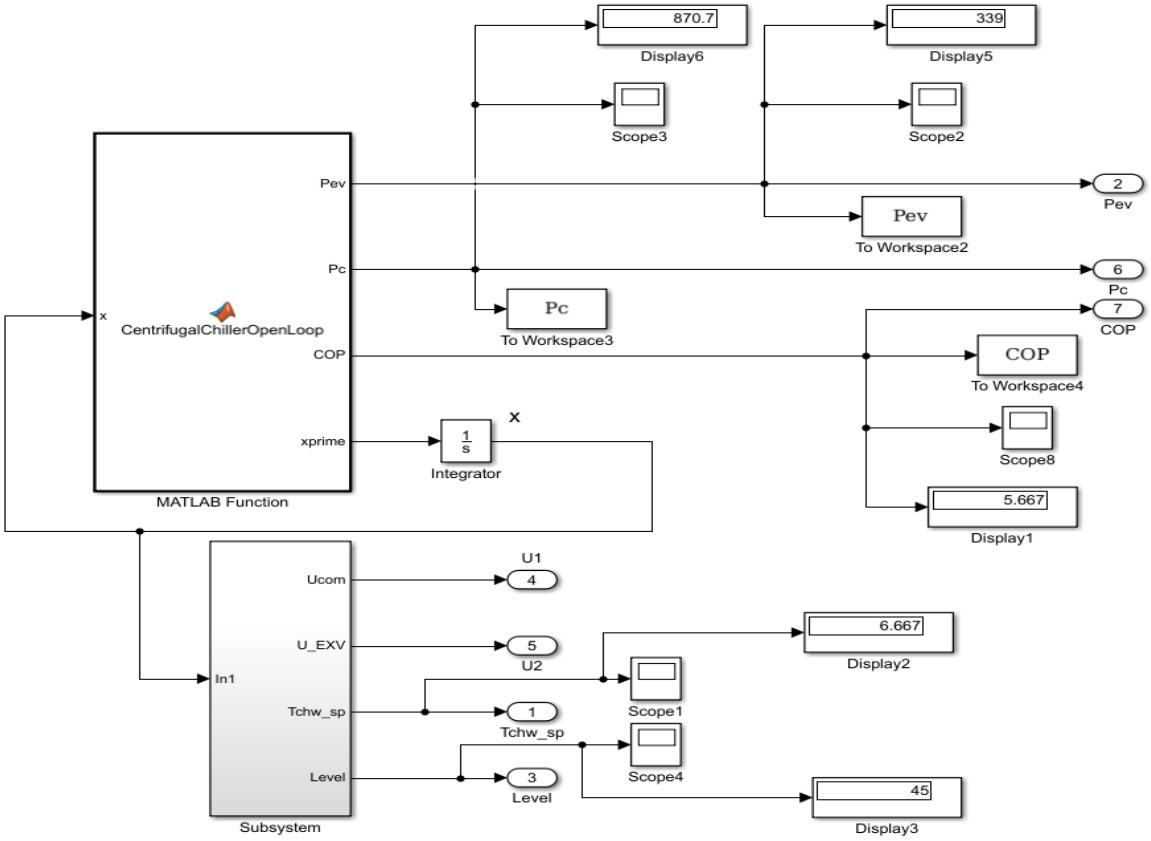


Figure 32. The SIMULINK model of chiller system in open-loop

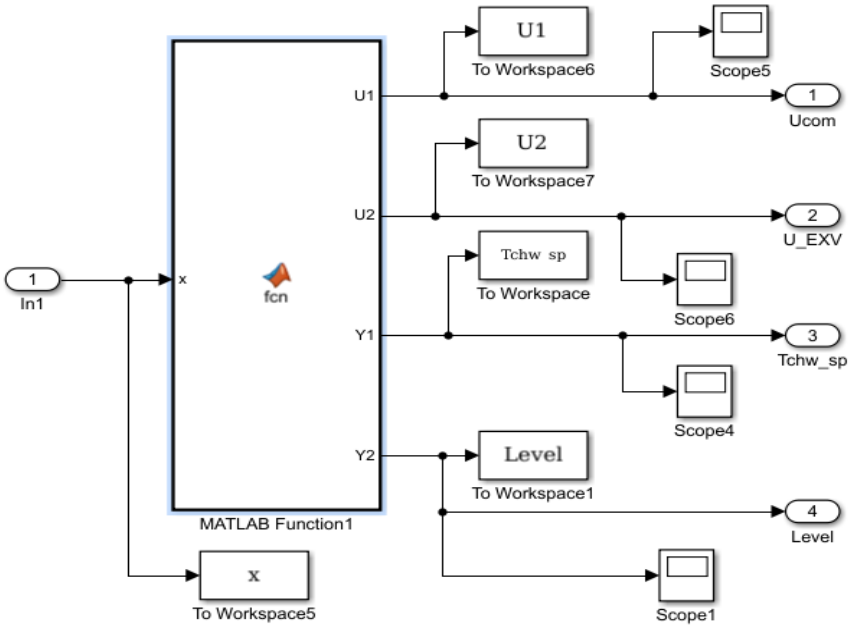


Figure 33. The SIMULINK block selector of chiller system inputs and outputs

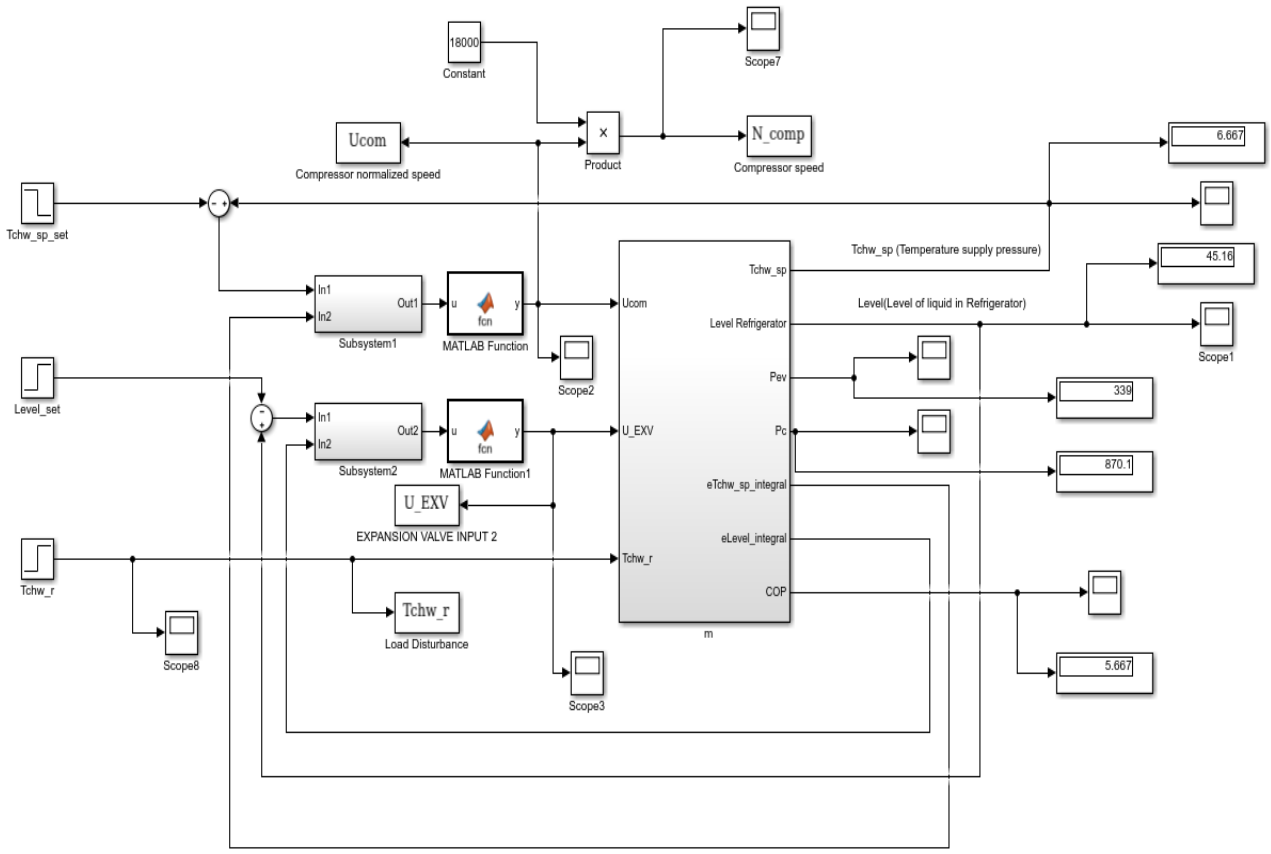


Figure 34. The closed-loop overall PI control of chiller system

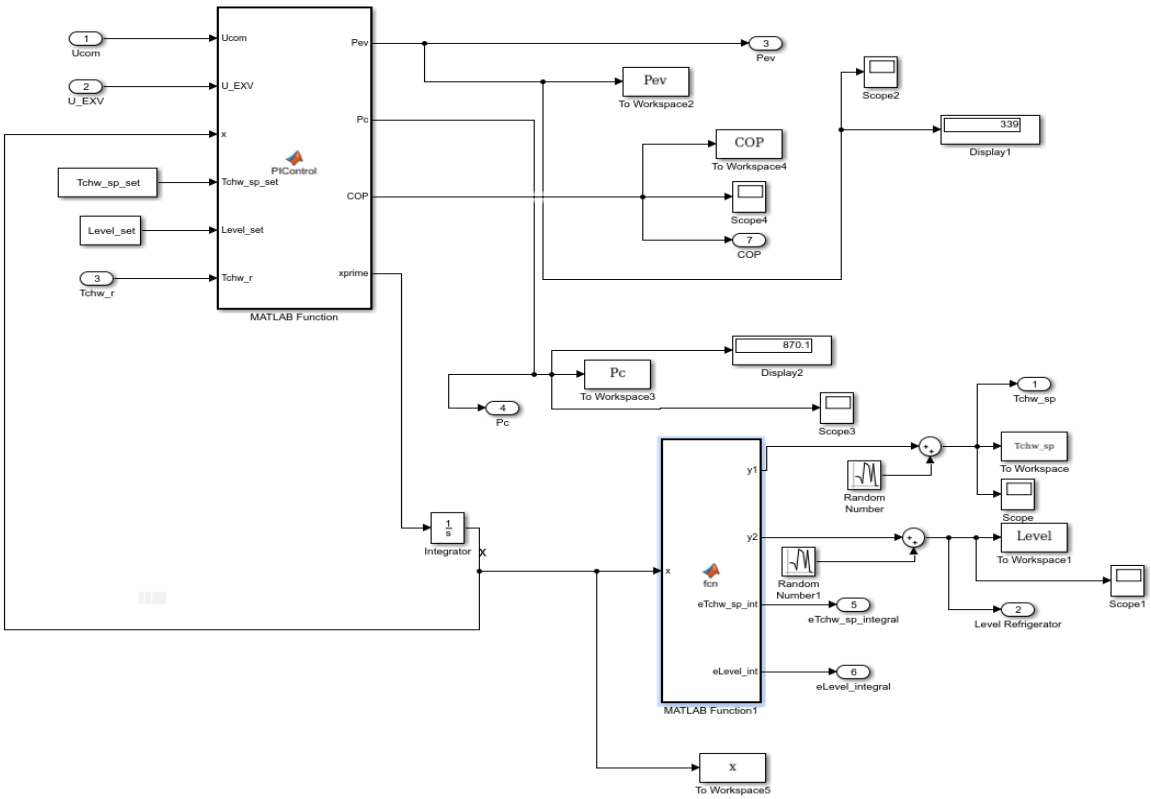


Figure 35. The SIMULINK block models of the state vector integrator and output selector

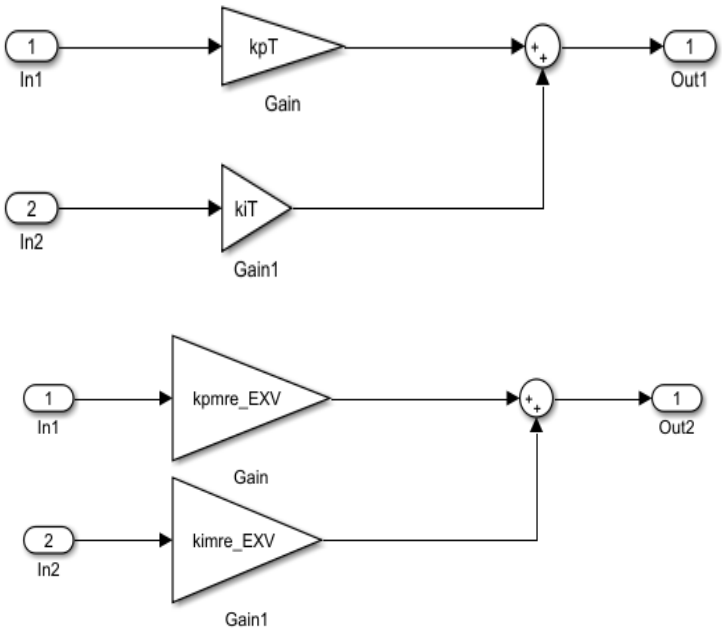


Figure 36 SIMULINK block of PI controllers for chilled water temperature (left) and liquid Refrigerant level (right)

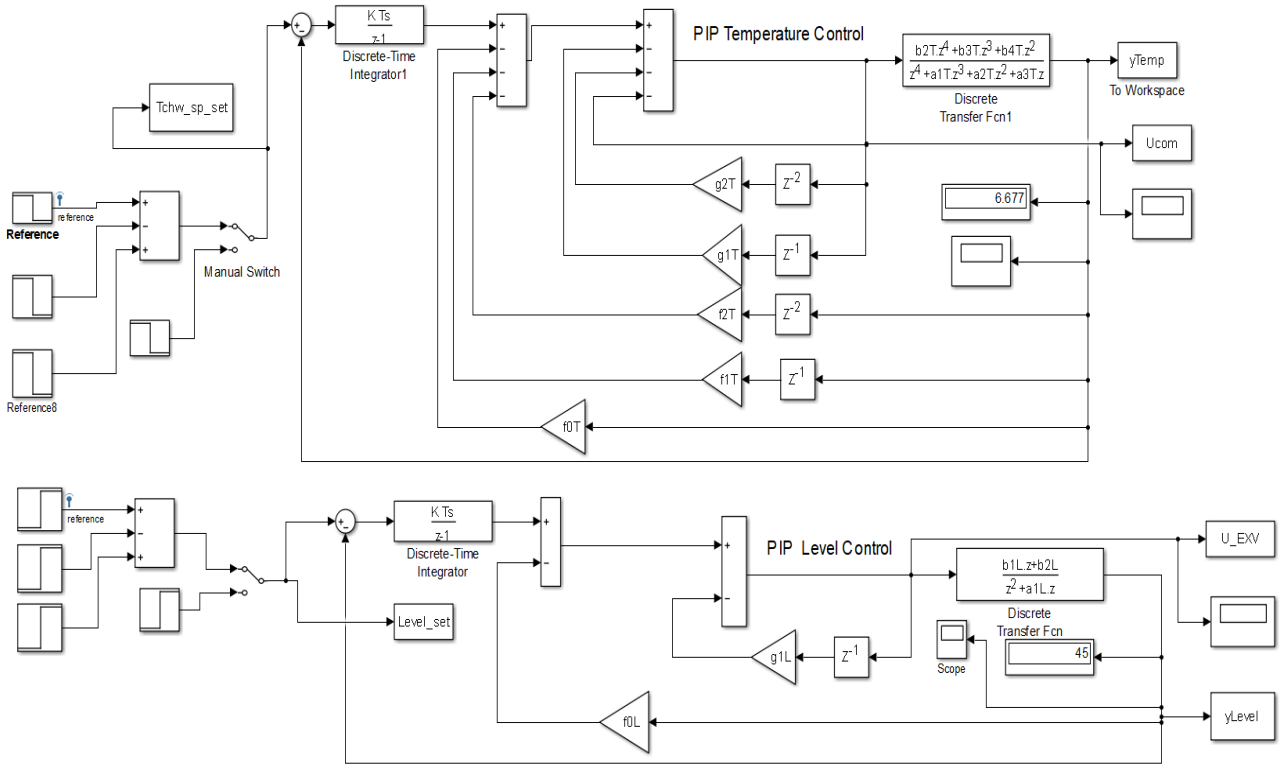


Figure 37. The PIP controller of chilled water temperature in Evaporator (top) and PIP controller of Liquid Refrigerant level in Condenser (right)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 284-294 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

The experience of implementation with Agile Business Process Management

Denis Rodríguez*, Enrique Silva Molina

Electrical-Electronic-Computer Faculty, National Engineer University, Nicaragua

ARTICLE INFO

Article history:
Received: 29 June, 2018
Accepted: 02 August, 2018
Online: 12 August, 2018

Keywords:
Agile Business Process Management
Business Process Management
Wiki

ABSTRACT

This paper is an extension of work originally presented at the Convention of Central America and Panama XXX VII- The Institute of Electrical and Electronic Engineers -2017. We have extended our previous work by presenting the initial results of how social technologies in particular wikis has the potential for Information Technology support of a cooperative community knowledge generation.
This paper outlines an experience in the implementation of a multidisciplinary research project – AGILe busIness PrOcess (AGILIPO), applied to an important organization in Nicaragua. In addition, it focused on improving key process by describing through collaborative tools which foster end user collaboration and organization knowledge.
We gathered evidence that blogs and podcasts couldn't afford to allow users on collaborative tasks of AGILIPO. The principal contributions of this work are the lessons learned from the experience of applying AGILIPO to an organization in Nicaragua, the key elements in achieving success in the application, the conditions that were brought together in order to use AGILIPO, the drawbacks during the experience and the relevant limitations of AGILIPO and how we can improve process by agile Business Process Management.

1. Introduction

Business Process Management (BPM) is a discipline which has been around since the early 90s, launched by the article by Hammer (1990) and reinforced by the book by Hammer and Champy (1993) on Business Process Reengineering (BPR). Hammer (1990) stressed that Information Technology (IT) made it possible for companies to undertake major revisions in the way they did work [1].

One of the methodologies of BPM agile found in the bibliographic searches carried out in this work was AGILIPO. In this literature review they were not experiences prior documented of the implementation of the AGILIPO methodology in an organization. Thus, what is innovative of this work is to implement AGILIPO in an organization in Nicaragua [2].

Agile BPM represents the next generation of business process management designed to flexibly address all types of processes to support all forms of work. It combines traditional Business Process IT-Management (BPM) style predefined processes, along with Adaptive Case Management (ACM) style dynamic work support. Agile BPM is designed to flexibly address all types of processes used to conduct business: structured, unstructured, and hybrid process types to support all forms of work [3].

The methodology used in researching this paper has a qualitative approach with a practical action-oriented research design, which the investigator carried out a literature review of the area of the knowledge of BPM and Agile BPM. It deepened through an analysis of the background of BPM and Agile BPM. The result of this analysis was the selection of a multidisciplinary research project AGILe busIness PrOcess (AGILIPO). Next, the investigator developed the experience to apply AGILIPO to an organization in Nicaragua. Based on this experience of applying AGILIPO, we learned a group of lessons [4].

This research project shows the lessons learned from the experience of applying AGILIPO to an organization in Nicaragua, the key elements to achieve success in the application, the conditions that we should have in order to use AGILIPO, the drawbacks during the experience and the relevant limitations of AGILIPO [4].

The principal contributions of this work are the lessons learned from the experience of applying AGILIPO to an organization in

*Corresponding Author: Denis Rodríguez, FEC-UNI, Managua, Nicaragua, +1202-974-3009, rodrigde@paho.org

Nicaragua. On the basis of this experience some limitations of AGILIPO were found [2].

From the experience of applying AGILIPO evidence it was learned that this increases the participation of the users in the phase of modeling and implementation of key processes. On the other hand, the search in the databases for research during this work showed the lack of research on the methodology agile BPM.

2. Analysis of BPM background

2.1. Definitions

What is a process?

A process corresponds to the representation of a set of actions (activities) that are done (carried out), under certain conditions (rules) and that can trigger or cause events [5].

What is a business process?

A business process is a set of activities, which promoted by events and carried out in a certain sequence create value for a client (internal or external) [5].

What is a key process?

A business process which, from management's point of view, is critical to customer service and satisfaction, has a competitive advantage, or has the success of the firm's strategy [2].

What is Business Process Management (BPM)?

Business Process Management (BPM) is a management discipline that integrates the strategy and goals of an organization with the expectation and needs of customer by focusing on end-to-end process. BPM comprises strategies, goals, culture, organizational structures, roles, policies, methodologies, and IT tools to (a) analyze, design, implement, control, and continuously improve end-to-end processes, and (b) to establish process governance [6].

After WWII, applying science to process became front and center as W. Edwards Deming and Joseph Juran taught the Japanese about the power of quality management. Their work and the work of others triggered a wave of Total Quality Management (TQM), spurred on by the publications of Deming and Juran in 1982 as shown below. The emphasis was not so much on the design of new processes, but on statistical measurements as a means of improving existing work practices and quality [7].

Then a decade later, the 1992 blockbuster books, Process Innovation and Reengineering the Corporation, hit corporate board rooms, and reengineering work through information technology took off. In this second wave of business process management, processes were manually reengineered, and through a one-time, big-bang activity, cast in concrete in the bowels of today's automated Enterprise Resource Planning (ERP) and other packaged systems. Although "downsizing" is the moniker most remembered from Business Process Reengineering (BPR), it was technological enablement—including office automation—that allowed companies to tear down internal silos and reengineer end-to-end business processes that spanned individual functional departments (silos) [7].

In the third wave of process management, the business process was freed from its concrete castings and made the central focus and basic building block of automation and business systems. Processes became first-class citizens in the world of automation. Change was the primary design goal because in the world of business process management: the ability to change is far more prized than the ability to create in the first place. It is through agile business process management that end-to-end processes can be monitored, continuously improved and optimized. Feedback of results, agility and adaptability are the bywords of the third wave [7].

Industry 4.0 denotes the Fourth Industrial Revolution, a term introduced in 2011 at the Hannover Messe, one of the world's largest trade shows, and since then widely used by German industry and government [8-5].

BPM as Support for Industry 4.0 and E-Commerce

The new industrial revolution will have a strong impact on the relation of BPM and e-Commerce because it moves manufacturing and production from a centralized to a decentralized paradigm. This will require a widespread adoption of smart interconnection of machinery and systems, not only at the same production site but also across the entire organizational ecosystem. An enormous potential opens up for innovation in business processes and in the way society interacts at a global level [9].

Agile BPM

One of the reasons why BPM is losing momentum is the lack of strategic alignment of its programs. On the one hand, organizations involved in BPM initiatives still mainly focus their efforts on the early stages of the BPM lifecycle, i.e. on process identification and discovery [10].

For many years, we have been arguing that traditional models of management, sponsorship, and project management have not coped with the increasing rate of change. In two separate Harvard Business Review articles, Donald Sull argues for a values-based agile model of business and Gary Hamel presents 25 grand challenges designed to redesign all aspects of management theory and practice to address the fact that "modern models of management" have reached their limits [11].

Business process management (BPM) encompasses the discovery, modelling, monitoring, analysis, and improvement of business processes. Traditional BPM limitations in addressing changes in business requirements have resulted in a number of agile BPM approaches that seek to accelerate the redesign of business process models [12].

Another alternative to highly engineered processes might he called "agile" methods. They are less focused on the specific steps to be followed in a process, and more oriented to the managerial and cultural context surrounding the process. Instead of detailed process flows, for example: agile methods might emphasize the size and composition of process teams, a highly iterative workflow, and a culture of urgency. This is the case, for example, in the agile method known as "extreme programming." [13].

Next-generation Agile BPM is designed to address all of the requirements of managing work in today's enterprise: from streamlining routine, repeated business processes to managing

dynamically evolving business cases involving teamwork, collaboration, and judgment across and among diverse sets of process participants [3].

Achieving successful BPM solutions starts with an agile approach. An agile approach supporting stronger business and IT collaboration on BPM and minimizes the challenges of ineffective requirement definition. By engaging business in the actual development process, design problems can be surfaced more quickly and this helps to reduce the typically long development cycles often seen in BPM. Finally, and perhaps most importantly, an agile BPM approach can institutionalize a partnership between business and IT by providing a foundation for shared commitment, role management and process ownership [14].

AGILe busIness PrOcess (AGILIPO)

AGILIPO - AGILe busIness PrOcess - is a multidisciplinary research project that has been put forward within the scope of a wide-raging research programme aimed at integrating the development of computer-based artifacts for organizations with the accumulated knowledge on organizational design [15].

AGILIPO approach is based on the following steps:

1. Defining the process according to its goals.
2. Defining the process according to its agency and context, using a classificatory framework based on organizational routines.
3. Describing the process using a wiki-like approach, where collaboration, user empowerment and tacit knowledge usage are key principles.
4. Fine-tuning the description through an ontological approach known as folksonomy.
5. Validation of the process.

AGILIPO approach steps are shown in Figure 1.

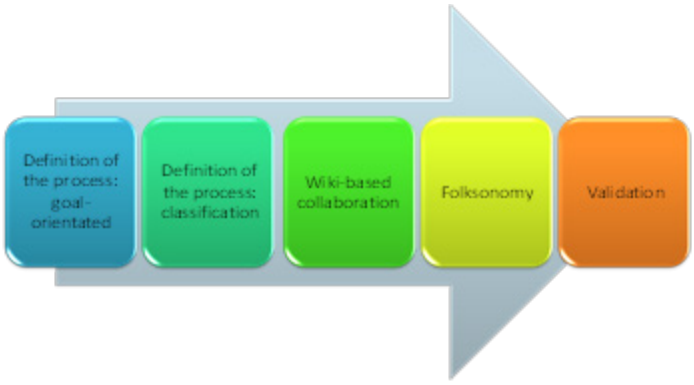


Figure 1: Proposal AGILIPO approach

Definition of the process: goal-oriented

Goal-orientation mirrors the approach we take to goals in our own lives and lends itself to business user involvement in the creation and management of processes. This also extends to routine tracking of plan execution to detect problems as they occur, or even better before they do, in order to take timely and appropriate actions [15].

Definition of the process: classification

On the topic of standardization (and eventual reuse) if we know the process type(s) and the process instance(s) that we are dealing with, this will save much time and effort at the outset. Classification of processes is a method which has been talked about in the literature. There have been many initiatives aimed at cataloguing generic business processes, each proposing classifications of their own, including the MIT process handbook (Malone, 1999) or the Process Classification Framework by the American Productivity and Quality Center's International Benchmarking Clearinghouse (APQC, 2006) [15].

Wiki-based collaboration

In the context of an agile BPM methodology, we proposed that a wiki-type tool can be created for the collective description of business processes. In this context, three aspects may be considered when evaluating wiki-type scenarios [15]:

1. The degree of organization of the BPM team.
2. The degree of specificity of wiki objects (goals, sub-goals, activities, roles, etc.).
3. The degree of desired process completeness.

Wikis, blogs/photoblogs and podcasts (and its video incarnation, the vodcast) carry the potential of complementing, improving and adding new collaborative dimensions to the many Web-based medical/health education, Continuing Professional Development (CPD), and research services currently in existence. They offer many unique and powerful information sharing and collaboration features. They also afford users the added advantage of reducing the technical skill required to use these features, by allowing users to focus on the information and collaborative tasks themselves with few delivery obstacles [16].

Folksonomy

Folksonomies are an emergent phenomenon of the social Web. They arise from data about how people associate terms with content that they generate, share, or consume (Gruber, 2006). It is claimed that Folksonomies have many advantages over controlled vocabularies or formal taxonomies. Tagging has dramatically lower costs because there are no complicated, hierarchically organized nomenclatures to learn. Users simply create and apply tags on the fly. Folksonomies are inherently open-ended and therefore respond quickly to changes and innovations in the way users categorize content (Wu et al., 2006) [15].

Validation

The modeling of business processes is never finished because the process itself is never complete. In order to overcome such a realization, a validation step must be adopted (Kuhne, 2008). Validation should give immediate and continuous feedback to business process designers about weaknesses and inconsistencies in possibly incomplete models. The established modeling process with sequential modeling, validation and evolution stages should be replaced by a modeling process with integrated validation support [15].

This research is not an attempt to generalize probabilistically the results to broader populations, nor necessarily obtain representative samples. This approach used interpretive practices that make the organization visible where research was developed and proposals made to transform the way in which people interact with the key processes. This became a representation in the form

of observations for the AGILIPO proposed methodology, academic articles, and diagrams of modeled key processes [2].

3. The experience of implementing AGILIPO in an organization in Nicaragua

Current approaches to BPM still work on the AS-IS/TO-BE paradigm, inherited from the Business Process Reengineering (BPR) era from the nineties. BPR is a top-down, holistic, and cross- cutting approach that takes months of analysis and impact assessment to achieve. [17-18]. The problems with the AS- IS/TO-BE approaches are related to the temporal gap between the modeling and implementation phases as well as the lack of involvement of the users [15]. These problems have been little addressed by BPM and agile BPM. That is why they are important at present.

After searching the databases of Scopues–Elsevier, ScienceDirect, Google Scholar and IEEE, articles on neither the implementation of AGILIPO nor experiences of AGILIPO in Nicaragua were found. Thus, what is innovative in this work is to implement AGILIPO in an organization in Nicaragua with live business processes.

Although Business Process Management implementation is usually associated with large scale business, implementing it on small medium enterprise may also be a good consideration for Small Medium Enterprises (SMEs) owners. SMEs are non-subsidiary, independent firms which employ fewer than a given number of employees [19]. Upon regarding this organization in Nicaragua as a SMEs, the implementation of AGILIPO should be taken into account.

The experience was carried out in the natural and daily environment of the PAHO/WHO Representative Office in Nicaragua in 2015 and included the key processes for the biennium 2014-2015 [2].

The key processes were:

1. Local and international purchases (procurements of goods).
2. Service contracts (contracts of providers).
3. Temporary advisor contracts (contract of temporary people).

Furthermore, this experience responds to the AGILIPO team requirement to improve even more the key processes of PAHO/WHO Nicaragua of the biennium 2016-2017. Besides, it tries to respond what types of other social network tools as blog and podcast could be feasible for AGILIPO.

The experience was divided into three stages:

The first one included semi-structured interview of project's managers and administrative-technical committee (operational structure linked with the review of processes in the Representative Office) [2].

The second stage included development of the experience using the proposal methodology AGILIPO with the key processes. For this activity there was created the AGILIPO team, which was integrated by a representative of each level of the organization chart of the Representative Office and the investigators [2].

The third one included an evaluation of the use of other social network tools (blogs and podcasts) for step 3 to 5 of AGILIPO and improves the key process of PAHO/WHO Nicaragua of the biennium 2016-2017 by using the best practices of modeling with Bizagi.

3.1. Results obtained in the first stage of the experience

The purpose of the first interview of the four project managers was to compile information on the knowledge of the interviewees of the methodology of process management that the organization has used in the last 6 years. In the interview all the interviewees agreed that the organization had tried to be efficient and that the companies should be flexible and adapt to change [2].

The preliminary results of the first interview warn the AGILIPO team that the project managers had little knowledge about BPM and agile BPM. Thus the investigators trained the interviewees on the basic concepts of BPM and agile BPM. In the opinion of the investigators a condition for using AGILIPO should be training of the interviewees and the AGILIPO team on the basic concepts of BPM and agile BPM to improve the knowledge of the interviewees about BPM and agile BPM.

The second interview of the project managers and the administrative technical committee, through questionnaire No. 2, had the purpose of compiling information on the knowledge of the interviewees on the possibility of the use of methodology agile BPM for management of the processes of the Representative Office. 89% of the participants thought that the agile BPM could be used in the operational processes. 100% believed that increasing the participation of the users could help to improve the management. 89% believed that it was feasible to use collaborative tools that exist in the organization [2].

The AGILIPO team agreed with the interviewees who were open to changes and improvements in the processes of the organization and in strengthening the use of tools already existing in the organization. The opinion of the interviewees were taken into account to help improving the process and to empower the users to share their ideas and proposals to work on the issues.

3.2. Results obtained in the second stage of the experience

In the second stage of the experience had participated five peoples of the administrative technical committee (75% of the committee) through a Wiki site [2].

The first round of the experience was carried out with the three operational processes of the general services unit. The definition of the process aimed at step 1 was complicated since the participants could not link the modeled processes with the goals of the Representative Office. The definition of the processes according to this agency (organization) and context was complicated because the participants did not understand the classification approach defined by AGILIPO and they did not reach a consensus. That issue did not allow participants to describe the process with the Wiki in conjunction of the folksonomy approach [2].

Based on the results of the first round; the Representative took a decision on how to model the key process of the WHO/PAHO Nicaragua. Below are the results of the execution of the key

processes of procurements of goods shown using the five steps of AGILIPO [2].

Step 1: Definition of the process: goal-oriented

In this step 75% of the interviewees did not link the process of procurement of goods to any product of its corresponding project in the Plan of the bi-annual Budget 2014-2015. Furthermore 100% of the interviewees did not define the objectives of the process. 75% of the interviewees did not reach a consensus on who should do the tasks (human or technological actors).

The definition of the process was guided by AGILIPO approach; it helped to detect the inconsistency that existed in the products of the projects of the biennium 2014-2015, which were not linked clearly to the key processes of the organization [2].

Step 2: Definition of the process: classification

Defining the process according to its agency and context, using a classificatory framework based on organizational routines [15].

Due to the difficulties that the members of the technical committee–administrator had in using the classification of step 2 of AGILIPO, the investigators decided to obtain evidence of this difficulty applying the third interview through questionnaire No. 3. The purpose of this questionnaire was to compile information on how they used the classification based on Howard-Grenville (2006) [4].

The questions of the questionnaire No. 3 were coded in the Figure 2 with this syntax: Questionnaire number X and Question number Y. Example: Q3Q1.

The questions of the questionnaire No. 3 for the process of procurements of goods are listed below:

1. Do you believe that the process of procurements is necessary? [Q3Q1]
2. Do you believe that the process of procurements has a clear definition? [Q3Q2]
3. You believe that the process of procurements has a clear definition of who should do it? [Q3Q3]
4. Do you believe that the process of procurements has a weak embeddedness ? [Q3Q4]
5. Do you believe that the process of procurements is flexible? [Q3Q5]
6. Do you believe that the process of procurements has some change probabilities? [Q3Q6]

Regarding the classification based on Howard-Grenville (2006), all the participants agreed on what had to be made. Although 75% believed that the process was not flexible, and they did not reach a consensus on who should make this process. Using this classification was complicated for the interviewees, due to a lack of understanding that the columns were related. For example, the embeddedness of the process, with the actors primary orientation, the performances of the flexible process, the changes in process over time, process labels and characteristics over time[2]. The overall of the answers of the questionnaire No.3 is shown in Figure 2.

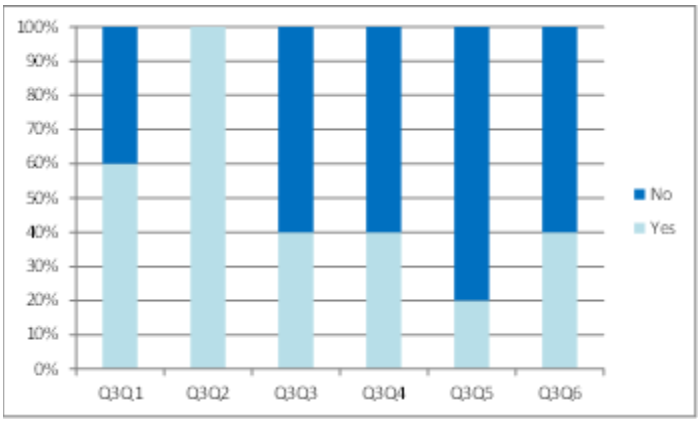


Figure 2: Abstract of responses to the questions of the questionnaire No.3

The classification of processes constituted a limitation of AGILIPO. Since all the participants did not understand how to use it. The author suggests changing to the MIT process classification or the framework of classification of processes of APQC [2].

Step 3: Wiki-based collaboration

Social technologies provide a range of collaboration and communication tools such as blogs, wikis, forums, chat platforms, etc. that support user interaction through social computing features. These features enable users to easily capture and share the knowledge and expertise that is needed to do their work. This sharing of information encourages collaboration, improves innovation, and targets relevant content to the people who have to see it [20].

Traditional process elicitation methods are expensive and time consuming. Recently, a trend toward collaborative, user-centric, on-line business process modeling can be observed [21].

Enterprise social networks are gaining momentum as a platform for collaboration between members of an enterprise, leading to the notion of Enterprise 2.0 [22].

Upon considering the issues suggested by the authors for the use of the Wiki:

- The level of organization of the team AGILIPO, in this case is made up of the investigator and the members of the technical committee–administrator. The mapping and description of the process, including the number of participants of different organizational units, were based on the rules established by the Representative Office in order to guarantee their consistency. The Wiki have a mechanics of horizontal work that provided the proposals of the participants in the description of the process.
- The degree of specificity of the objects of the Wiki (goals, activities, roles). The structure of the data, related to the definition of the process was based on the knowledge of the participants in the experience maintaining the greatest possible simplicity.
- From the standpoint of social intelligence, we can observe that the more collective intelligence is needed for a given task, a lesser degree of organization is required. . In other words to greater participation in the development of an artifact, the greater the probability of its completion is increased. During

the execution of the experience, a description by consensus was obtained for the most part.

Wiki was used to model the key processes, through the cooperative knowledge of the AGILIPO team. With the Wiki an active participation of the AGILIPO team in the modeling of the key process was seen. This participation made it possible to carry out contributions to the modeled key process [2].

Step 4: Folksonomy

The folksonomy was the most interesting step in the experience. Since the participants were accustomed to working with methodologies of structured management as the logical Framework and the Results-based Management. These methodologies are from the top down. In the opinion of the participants, this made it difficult for them to be able to prepare the labels without a hierarchy resulting in the participants having equal decision-making power [2].

The transcription of the Wiki in the creation of labels of the process of procurement of goods, through the folksonomy, was developed in four interactions. The last round is shown below [2]:

- Create request of purchase
- Confirm if it is planned
- Authorize request
- Notify It staff member the change
- Generate offers
- Approve offer
- Notify there is not offer
- Review budget
- Notify change in budget
- Create purchase order
- Authorize purchase order
- Cancel purchase order
- Send purchase order to supplier
- Receive products
- Review products
- Notify claim of goods
- Pay invoice to supplier
- Close purchase order

The labels created by the members of the AGILIPO team are tabulated in Table 1. These labels described the process of procurement of goods (Process 1). These labels were grouped by synonyms synthesizing them in verbs or actions that describe the analyzed process. Finally, the labels are ordered by coincidences in order which they were written by the members of the AGILIPO team and the results of the tasks of the process analyzed [2].

Table 1: Review of labels of the folksonomy – Process 1

Labels	Person 1	Person 2	Person 3	Person 4	Person 5
Create	1				
Request		1	1		
Request purchase				1	
Authorize request		1			1
Quotations		1			1
Authorize quotation	1				
Create purchase			1		1
Notify vendor		1		1	
Reject request			1	1	
Receive goods	1				1
Check goods			1	1	
Goods claim		1		1	
Business rules ok	1				
Pay to vendor				1	1
No money	1			1	1

Finally, the labels are ordered by coincidences and in order in which they were written by the members of the AGILIPO team and the results of the tasks of the process analyzed [2].

- Register
- Request
- Request purchase
- Authorize request
- Offers
- Approve offer
- Create purchase
- Notify supplier
- Reject request
- Receive goods
- Review goods
- Demand of purchase
- Business rules OK
- Pay supplier
- Defund

The folksonomy used jointly with the Wiki was the reason the participants described the key processes and created the labels that helped to improve the key processes [2].

Step 5: Validation

Taking into consideration that PAHO/WHO Nicaragua cannot implement changes in their key processes as part of this work. The

investigators took a decision of carry out a simulation in order to validate the modeled processes [2].

The procurement units had six kinds of purchases services. As part of the validation, the investigators reviewed those six cross functional flowchart and create one. Based on that cross functional flowchart, the investigators create the old BPM diagram of procurements process. The old BPM diagram of procurements is shown in Figure 3.

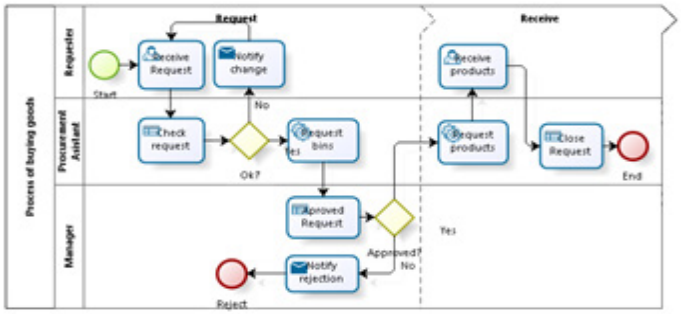


Figure 3: Old BPM diagram of procurements process

The procurement units had two kinds of services contracts. As part of the validation, the investigators reviewed those tow cross functional flowchart and create one. Based on that cross functional flowchart, the investigators create the old BPM diagram of services contracts. The old BPM diagram of services contracts is shown in Figure 4.

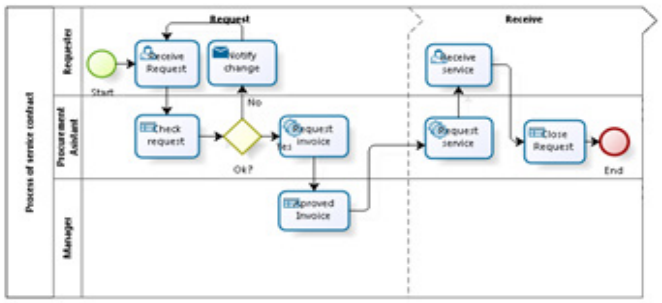


Figure 4: Old BPM diagram of services process

The Human Resources (HR) units had three kinds of services contracts. As part of the validation, the investigators reviewed those three cross functional flowchart and create one. Based on that cross functional flowchart, the investigators create the old BPM diagram of recruitment temporary staff. The old BPM diagram of temporary staff is shown in Figure 5.

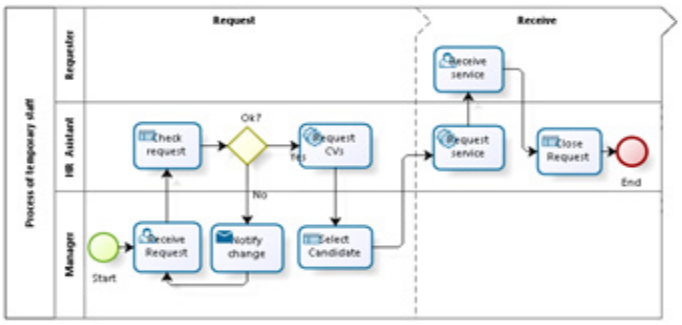


Figure 5: Old BPM diagram of temporary staff

Based on the log of the Wiki in conjunction with the labels produce through the folksonomy process we create a new BPM diagram of the process of procurement of goods, services contracts, and recruitment of temporary staff.

The resulting new BPM diagram of procurements process based on the best practices in modeling of Bizagi is shown in Figure 6. The improvements of the old BPM diagram were the following:

- Use the standard BPMN (Do not use unit names).
- Simplification of the diagrams (Combine two roles which bellow to the same organizational unit).
- Maintain the logical sequence and cleaning (Eliminate duplicated end events).

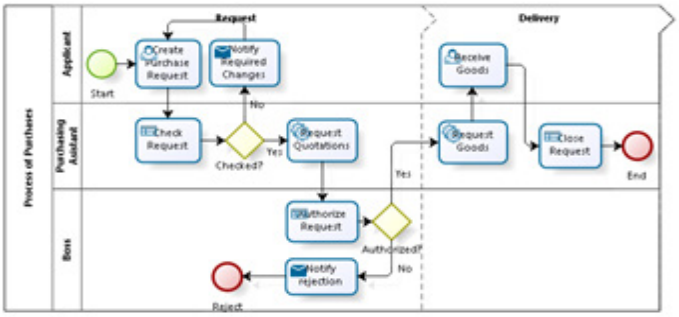


Figure 6: New BPM diagram of procurements process

The resulting new BPM diagram of service contracts process based on the best practices in modeling of Bizagi is shown in Figure 7. The improvements of the old BPM diagram were the following:

- Use the standard BPMN (Use standard job title).
- Maintain the logical sequence and cleaning (Add rejection task after authorize tasks).

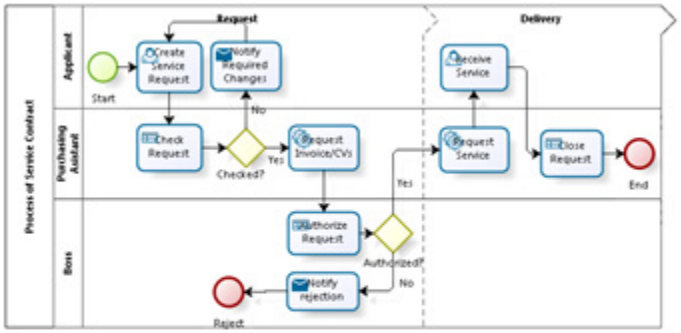


Figure 7: New BPM diagram of service contracts process

The resulting new BPM diagram of recruitment temporary staff process based on the best practices in modeling of Bizagi is shown in Figure 8. The improvements of the old BPM diagram were the following:

- Simplification of the diagrams (Combine two roles which bellow to the same organizational unit).
- Maintain the logical sequence and cleaning (Transfer two tasks from the boss to the applicant, Add a rejection task after a decision task).

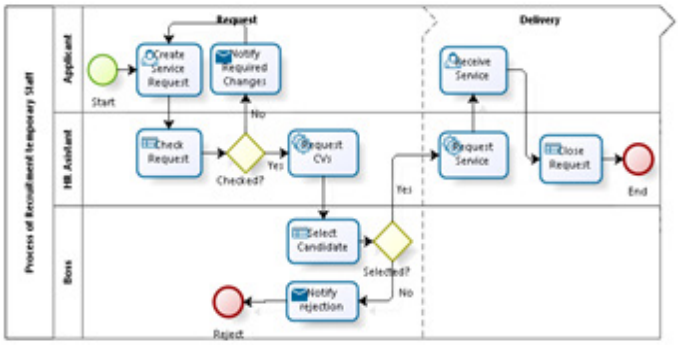


Figure 8: New BPM diagram of recruitment temporary staff process

Considering the limitations of time on the work of the thesis and the regulations of PAHO/WHO to authorize changes in their processes, a validation was carried out through a simulation on the basis of a scenario: 30 calendar days of execution, along with the personnel that worked in the Representative Office during the development of this research [2].

Bizagi Simulation comprises of four levels. Each subsequent level incorporates additional information exhibiting more complexity than the preceding one, thereby providing a detailed analysis of your processes [23].

1. Process Validation
2. Time Analysis
3. Resource Analysis
4. Calendar Analysis

To illustrate each of the simulation levels let us consider the level 1 for the Procurement Process.

Level 1: Process Validation

This level of the simulation validates gateways, messages, decisions and routing and all the sequence flow are fine. The results of the Bizagi simulation level one for Procurement Process are tabulated in Table 2.

Table 2.Results of process validation of Procurement Process

Name	Type	Tokens completed
Process of Purchases	Process	1000
Start	Start event	1000
Checked?	Gateway	2022
Authorized?	Gateway	1000
Reject	End event	504
End	End event	496
Create Purchase Request	Task	2022
Check Request	Task	2022
Request Quotations	Task	1000
Notify Required Changes	Task	1022
Authorize Request	Task	1000
Notify rejection	Task	504
Request Goods	Task	496
Receive Goods	Task	496
Close Request	Task	496
Process of Purchases	Process	1000

3.3. Results obtained in the third stage of the experience

In the third stage of the experience participated five peoples of the administrative technical committee (75% of the committee) through a Google's blogger, WordPress site with podcast, a SharePoint blogs site with podcasts, and a SharePoint blogs site.

The AGILIPO team tested the Google's blogger site. Only 20% of the AGILIPO team created his/her account. Besides, the team tested Apple's Podcasts. The team was unable to subscribe to a podcast channel on iTunes. The team agreed that they preferred a technology available in his working environment. Microsoft SharePoint is a technology that includes blogs and RSS feeder and is available on PAHO/WHO.

As part of evaluate the use of podcast and blog for step 3 to 5 of AGILIPO. The AGILIPO team used the outcome of the step 1 and 2 in stage two of this experience.

Step 3: Podcast-based collaboration

The investigators created a SharePoint blogs site with RSS feeds and configure the Outlook RSS reader watching the SharePoint site. After that, they made a quick guide to help AGILIPO team to create a short podcast and upload the audio files on the SharePoint site. Only 20% of the team could create three small audio files of one or two minutes. To deal with this issue the investigators trained to AGILIPO team how to create an audio file. After that, only 40% of the team could create three additional small audio files of two minutes. All the AGILIPO team agreed that describing the process using podcast approach was not easy to follow the discussion and describe the process. They spent much time listening all audio files.

Step 4: Folksonomy through Podcast

The AGILIPO team used the SharePoint blogs site to tune the description through folksonomy. Only 40 % of the AGILIPO team could create eight audio files for same number of labels. The AGILIPO team agreed that tune the description by folksonomy through Podcast was very confused. They said that was so complicated listened audio files and understand the labels.

Step 3: Blogs-based collaboration

The investigators created a SharePoint blogs and a quick reference guide to help AGILIPO team to post information on the SharePoint site. The investigators started the blogs and all AGILIPO team members posted information to describing the key process of PAHO/WHO Nicaragua. All AGILIPO team agreed that they felt a little concern about who posted which information, rejected or support some information. They preferred the Wiki approach, because they do not see audit tracks and felt more comfortable given feedbacks.

Step 4: Folksonomy through Blogs

In the case of fine-tuning the description of the process through folksonomy the AGILIPO team agreed that was more complicated to follow the comments, feed backs and get a consensus. They spend more time reading the blogs to label the key process of PAHO/WHO Nicaragua. The AGILIPO team agreed that the Wiki approach was easier to describing the process and tune the description through folksonomy.

The transcription of the SharePoint blogs in the creation of labels of the process of service contracts through the folksonomy, was developed in four interactions. The last round is shown below

- Create request of contract
- Authorize request
- Notify changes
- Review budget
- Notify change in budget
- Authorize contract of service
- Cancel contract of service
- Send to review in department of legal advisory services
- Sign contract
- Request payment
- Pay invoice to provider
- Provider offers services
- Request report of services
- Pay the bill
- Close contract of service

The labels created by the members of the AGILIPO team are tabulated in Table 2. These labels described the process of services (Process 2). This process combines the process of services and recruitment of temporary staff and this is an improvement of both services. These labels were grouped by synonyms synthesizing them in verbs or actions that describe the analyzed process. Finally, the labels are ordered by coincidences in order which they were written by the members of the AGILIPO team and the results of the tasks of the process analyzed.

Table 3: Review of labels of the folksonomy – Service 2

Labels	Person 1	Person 2	Person 3	Person 4	Person 5
Create	1			1	1
Check		1	1		
Request invoice		1		1	
Authorize request	1				
Request service				1	1
Authorize Invoice	1	1			
Notify vendor					1
Notify changes		1		1	1
Receive services	1			1	
Pay the bill	1				
Close request		1	1	1	
Service claim				1	
No money			1		

Finally, the labels are ordered by coincidences and in order in which they were written by the members of the AGILIPO team and the results of the tasks for the improve process of services are:

- Create
- Check
- Request invoice
- Request service
- Authorize invoice
- Notify changes
- Receive services
- Close request

Improve the key process of PAHO/WHO Nicaragua 2016-2017

After analyzing the description of the improved process of service contracts using blogs approach and applying the best practices of Bizagi, the investigators create a new BPM diagram of the improved process of service contract. The AGILIPO team agreed that some steps of both process could be combined and improve the logic sequence of the process. The result diagram was easy to understand for the team.

The resulting new BPM diagram of procurements process based on the best practices in modeling of Bizagi is shown in Figure 9. The improvements of the old BPM diagram were the following:

- Simplification of the diagrams (Combine two roles which bellow to the same organizational unit).
- Maintain the logical sequence and cleaning (Eliminate duplicated tasks).

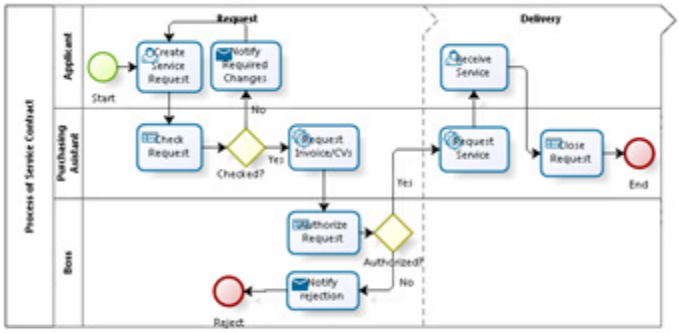


Figure 9: New BPM diagram of the improved process of service

4. Learning Lessons of the experience

The principal contribution of this research project carried out and synthesized in this article is described as a group of lessons learned about the basis of the experience to apply AGILIPO to an organization in Nicaragua. The lessons learned are detailed below [2]:

- Based on the experience of applying AGILIPO, we learned that a key element to achieve success is to write questions in order to determine the key processes.

- It was learned that a condition to use AGILIPO should be the training about BPM concepts for the members of the AGILIPO team..
- It was found that another condition to use AGILIPO should be the definition of the level of training, roles, and level of the organization chart of the participants who will be part of the AGILIPO team.
- It was noticed that the AGILIPO team had several drawbacks in classifying the processes; it was very difficult for the participants to use of AGILIPO classification.

In the second stage of the experience we modeled no key process and the AGILIPO team realized that the outcomes of this approach did not have sense. After that, the manager of this organization recommended model key process. The AGILIPO methodology does not provide any information about it and based on this experience, we learned that we should model key process with this methodology.

The experience in applying AGILIPO provided evidence of the conditions necessary to use AGILIPO. One is the training of the basic concepts of BPM and BPM of the AGILIPO team [2]. During the first stage of the experience the investigators interviewed project managers and realized that they had few knowledge about BPM and agile BPM. The investigators suggest to the AGILIPO team to have a training session for all interviewers and the AGILIPO team.

The experience to apply AGILIPO gave evidence on the conditions in order to use AGILIPO. One is to define the roles and the levels of the organization chart that they should be represented in the equipment AGILIPO [2]. Based on experience the investigators suggested including all levels of the organization chart and key persons with a deep knowledge of the unit and organization.

In addition, the experience in applying AGILIPO, there is evidence of some limitations in the application of AGILIPO, which are the classification of the processes. Thus, the investigators suggest change to the classification of processes of the Massachusetts Institute of Technology (MIT) or the framework of classification of the processes of the American Productivity and Quality Center. (APQC) [2].

The experience to apply AGILIPO gave evidence on the issues of working with podcasts to implement AGILIPO. The AGILIPO team agreed that was hard to follow a description of a process listening podcasts. Besides, tune the description by folksonomy. In the case of using blogs to describe and fine-tuning the description of the process through folksonomy the AGILIPO team agreed that was more complicated to follow the comments, feed backs and get a consensus. They spend more time reading the blogs to label the key process of PAHO/WHO Nicaragua.

The experience in applying AGILIPO, there is evidence of some limitations of podcasts and blog and the Wiki is feasible for AGILIPO.

In this paper it was verified that the use of social network tools is useful in the management of business processes. Since these can facilitate the interaction within the organization [24].

In addition, the experience in applying AGILIPO and responding a request of the AGLIPO team for improve the process obtained in the second stage of the experience. The improved process of services contracts combined effectively the process of recruitment temporary staff and services contracts. The result diagram was very easy to understand and few tasks.

After analyzing the description of the improved process of service contracts using blogs approach and applying the best practices of Bizagi, the investigators create a new BPM diagram of the improved process of service contract. The AGILIPO team agreed that some steps of both process could be combined and improve the logic sequence of the process. The result diagram was easy to understand for the team. In 2016 PAHO/WHO started the second phase of the implementation of the new Enterprise Resources Planning (ERP) and that ERP reduced most of the process.

The authors hope that other research of the proposal of the AGILIPO methodology is developed in other organizations using the Wiki.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank Dr. Socorro Gross, Representative of PAHO/WHO Nicaragua to permit use information on the Representative Office to conduct this research.

References

[1] M. Hammer, Reengineering Work: Don't automate, obliterate, Harvard Business Review, vol. 68, 1990.
[2] D. Rodríguez, "Gestión por Procesos de Negocio (BPM) – Una experiencia de aplicación de AGILIPO en una organización en Nicaragua", 2017.
[3] T. Koulopoulos, "Taming the Unpredictable". Future Strategies Inc. 1st Edition.
[4] D. Rodriguez, E. Silva, The experience of implementation with Agile Business Process, Management, CONCAPAN XXX –VII-IEEE, 2017.
[5] B. Hitpass, "BPM: Business Process Management - Fundamentos y Conceptos de Implementación". BHH Ltda. Cuarta edición. 2017.
[6] T. Benedict, N. Bilodeau, P. Vitkus, E. Powell, D. Morris, M. Scarsig, D. Lee, G. Field, T. Lohr, R. Saxena et al., "BPM CBOK Version 3.0: Guide to the Business Process Management Common Body Of Knowledge version 3.0", 3nd Edition, CreateSpace Independent Publishing Platform,2013.
[7] J. Sinur, J. Odell, P. Fingar. Business Process Management: The Next Wave, 2013.
[8] Germany Trade and Invest (GTAI), Promoted by Federal Ministry for Economic Affairs and Energy in Accordance with a German Parlament Resolution. Berlin, 2016.
[9] B. Hitpass, H. Astudilo, Industry 4.0 Challenges for Business Process Management and Electronic-Commerce, Journal of Theoretical and Applied Electronic Commerce Research, Universidad de Talca, Chile, 2019.
[10] M. La Rosa, "Strategic Business Process Management", Queensland University of Technology, 2016.
[11] R. Thomsett, "Agile Business: The Final Frontier", www.cutter.com. 2010.
[12] M. Zacarias, P. Ventura, "Agile Business Process and Practice Alignment Methodology: A Case-Study-Based Analysis".IGI Global, 2018..
[13] P. Bernus, J. Blazewiez, G. Schmidt, M. Shaw, "International Handbooks on Information Systems", 2010.
[14] S. Simmons, Steele, M., "BPM Voices: Synchronicity: An agile approach to business process management". 2012.
[15] R. Meziani, R. Magalhaes, "Proposal for an Agile Business Process Management Methodology". First International Workshop on Organizational Design and Engineering. 2009.
[16] M. Kamel, I. Maramba, S. Wheeler, "Wikis, blogs, and podcasts: a new generation of Web-based tools virtual collaborative clinical practice and education", BMC Medical Education, 2006.

[17] T. Morgan, Business Rules and Information Systems: Aligning IT with Business Goals, Addison Wesley, 2002.
[18] M. Cumberlidge, "Business Process Management with JBoss, jBPM", Packt Publishing –Birmingham, 2012.
[19] S. Allibang, "BPM: Business Process Management (BPM) of Small and Medium Enterprises", DBest Reads, 2016.
[20] A. Kumar, Business Process Management, Taylor and Francis, 2018.
[21] F. Dengler, D. Vrandecic, "Wiki-Based Maturing Of Process Descriptions". International Conference on Business Process Management. 2011.
[22] O. Hatzi, G. Meletakis, M. Nikolaidou, D. Anagnostopoulos, "Collaborative management of applications in enterprise social networks". International Research Challenges in Information Science, 2014.
[23] Bizagi Process Modeler User Guide, 2016.
[24] M. Rosing, A. Scheer, H. Scheel, "The Complete Business Process Handbook: Body of Knowledge from Process Modeling to BPM", Volume I, Morgan Kaufmann Publisher, 2015.
[25] M. Zacarias, P. Ventura, "Agile Business Process and Practice Alignment Methodology: A Case-Study-Based Analysis".IGI Global, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 289-313 (2018)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

An Empirical Study of Icon Recognition in a Virtual Gallery Interface

Denise Ashe, Alan Eardley*, Bobbie Fletcher

School of Computing and Digital Technologies, Staffordshire University, ST4 2DE, U.K.

ARTICLE INFO

Article history:
Received: 13 September 2018
Accepted: 07 November 2018
Online: 29 November 2018

Keywords:
Interface usability
Icon recognition
Icon intuitiveness
Icon design

ABSTRACT

This paper reports on an empirical study (an extension of a pilot study) that analyses the design of icons in a German 3-D virtual art gallery interface. It evaluates the extent to which a sample of typical computer users from a range of ages, educational attainments and employments can interpret the meaning of icons from the virtual interface taken 'out of context' and 'in context'. The study assessed a sample of 21 icons representing the 'action', 'information' and 'navigation' functions of the virtual interface using a new Icon Recognition Testing method (IRT) developed by the researchers from existing usability test methods. The Icon Recognition Rate (IRR) of the icons was calculated and they were classified as 'identifiable', 'mediocre' or 'vague' in a novel and useful classification system. The IRT results show that the IRR of almost a quarter of the icons was below the 'identifiable' standard, which could seriously compromise the usability of a virtual interface. A comparison is made, using textual and thematic analysis, between the participants' understanding of the icons' meaning in and out of context and of the effect of positioning icons in relation to their virtual surroundings and of grouping them in tool bars. From the findings of the study, conclusions are drawn, and recommendations are made for economical icon redesign and replacement. It is suggested in the conclusions that further research is needed into how designers' conceptual models can be better matched to users' mental models in the design of virtual interfaces by bringing user profiles into the study.

1. Introduction

This paper is an extension of a pilot study by Ashe *et al.* [1] into the effectiveness of icon design in a virtual gallery interface that was presented in the e-Tourism stream of the International Conference in Information Management at Oxford University in May 2018 (ICIM2018). Experience with that pilot study and feedback from reviewers informed a fuller research project, which forms the extended work in this paper. The second part of the research project used a larger, more representative sample of participants, took account of the context within which the icons are understood and added textual and thematic analysis to the research. The present paper therefore contains more detail about the research methodology and the data analysis and its results, which will allow the research to be replicated. The pilot study examined a sample of virtual tours of museums and galleries, including the Smithsonian Natural History Museum in Washington, D.C. [2], the Louvre in Paris [3], Oxford University Museum of Natural History [4] and the portal Virtual Tours [5,] that currently includes more than 300 'Museums, exhibits, points of special interest and real-time journeys' [6]. The study showed that icons are an important part of this generation of virtual interfaces as the main way of performing interactive tasks such as navigation, initiating actions and obtaining information [7]. The virtual interface itself is a complex sign system [8] containing components (e.g. buttons, icons and scroll bars) through which the user interacts with the system [9]. The icons can be symbols, images or pictures [10] that communicate meaning [8] without textual description [11-12].

This provides icon-based interfaces with the potential to overcome language barriers [10, 13], which can be important in an international context such as a cultural attraction. Icons used as shortcuts to a function (e.g. a printing icon in a word processing package) should provide the user with a memory aid to increase his or her ability to recall and to recognize the intended function without needing further instructions [14-15]. Successful recognition depends on the user's familiarity with that type of interface and experience of using that icon [1] and greater familiarity and experience should therefore allow a more abstract (i.e. less concrete) icon to be used in the design of the interface.

*Corresponding Author: Alan Eardley, Email: w.a.eardley@staffs.ac.uk

Gatsou [15] cites the work of Nadin [16], who uses a calculator icon to demonstrate the principles of concreteness and abstraction, as shown in Figure 1.

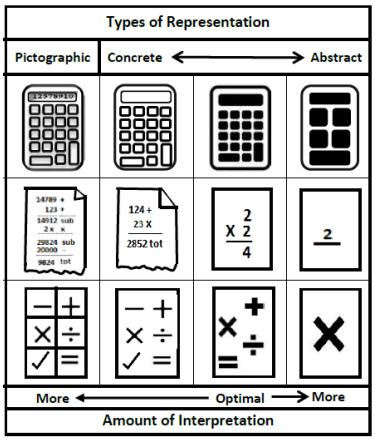


Figure 1: Types of icon representation
Adapted from Nadin [16] and Gatsou, *et al.,* [15]

Scalisi [17] suggests that users need an initial period of learning the interface to understand the icons through 'visual codification'. This may come easily with an office package that is used every day but may not be possible with a rarely-used interface such as a virtual gallery [1]. Icons may resemble to a greater or lesser extent the objects or functions that they represent [17] and the closeness of this relationship is the 'semantic distance', which is "*important in determining the success of icon usability*" [18].

Arnheim [19] discusses the relationship between 'concreteness and abstraction' stating that, "*Images can serve as pictures or as symbols; they can also be used as mere signs*", implying that increased user familiarity can allow an icon to be simplified yet still allow the user to understand it. The pilot study supported this view, suggesting that the closer the semantic distance, the more likely the users were to understand the icon's function and meaning. Conversely, the more abstract the icon (i.e. the greater its semantic distance) the more generally useful it could be in a variety of contexts, although correct recognition of the icon's meaning could be more difficult in a specific case.

For example, the icon for printing a document could be a photograph of the actual printer to which the file could be sent. That would promote easy recognition and could be useful, for instance to locate the correct printer in a room, but would involve having a different icon for every available model of printer. This would make it difficult for users to learn the general meaning of the 'printer' icon in other instances and applications.

2. Icon Usability

The pilot study [1] reviewed the literature on icon usability testing and took the definition of Ferreira *et al*. [20] who cite the work of Barr, *et al*. [14] stating that an icon is successful, "…if the interpretant of the user [i.e. the user's understanding] matches the object that the designer had intended with that sign, and [it is] unsuccessful otherwise" [20, p 2]. In other words, a recognizable (i.e. 'identifiable') icon should be easy to interpret and be unambiguous in order for it to succeed. This formed the baseline 'measure of success' used in the pilot study.

A range of different icon usability testing methods were reviewed in the pilot study, such as Icon Understandability Testing [21], [12], Test with Comparison [13], Matching Method [22], Icon Intuitiveness Testing [23] and Standard Usability Icon Testing [23]. From this review the Icon Intuitiveness Test (IIT) was selected for the study. The method was felt to be the most suitable as it seeks to find out how well users interpret and recognize icons using their existing insight and experience. Nielsen and Sano [23] describe a paper-based IIT as used by Sun Microsystems. Ferreira *et al*. [20] used a paper-based IIT and Foster [24] suggests that the IIT can be administered on a computer or on paper. Bhutar *et al*. [13] conducted a similar 'test without context' using an MS PowerPoint® presentation and paper-based questionnaires.

3. Pilot Study

Extending a previous study by Bhutar *et al.* [13], the modified IIT used in the pilot study adhered to the following guidelines:

- With one exception (i.e. Icon 1) the icons did not have text labels attached [23, 25] so their effectiveness relied entirely on their functioning as signs;
- The icons were not displayed in the actual interface (i.e. they were taken out of context), so the participants had no external visual cues to their meaning;
- Only one icon was made visible at a time so participants had no clues to their meaning from their sequence or by association.

Previous studies by Ferreira *et al*. [20] had used the standard ISO 9186:2001 benchmark [26] of 66% for successful icon recognition. Gatsou *et al*. [15] adopted the more stringent standard ISO 3864:1984 [27] which has a slightly higher benchmark, in which a success rate above 66.7 % was considered as 'good' and below that as 'low'. A similar scale by Howell & Fuchs [28], was adapted for use in the pilot study. With this scale, icons achieving 60% Icon Recognition Rate (IRR) or above are classed as 'identifiable', whereas icons scoring less than that are 'unsuccessful' in conveying their meaning.

The adaptation for the pilot study further divided these 'unsuccessful' icons into 'mediocre' (30% - 59% IRR) and 'vague' (0% - 29% IRR) as shown in Table 1.

Table 1: Icon Recognition Rates and Classifications

Icon recognition rate (IRR) classification	
IRR (%)	***Classification***
60 – 100	Identifiable
30 – 59	Mediocre
0 – 29	Vague

The research required as a subject an advanced interface containing icons that are capable of a number of different interpretations and which carry out defined functions. A virtual art gallery was felt to meet these requirements and a search on the World Wide Web identified more than 100 possible candidates. A German 3-D virtual art gallery was eventually selected for the test, as it was felt to be representative of its type [1]. For ethical reasons the site is referred to as 'Artweb.com'. The test examined the

users' understanding of the icons when taken 'out of context' (i.e. without reference to their use in the actual interface).

3.1. Icon Intuitiveness Test

All 21 icons used in the pilot study IIT were selected from the 'Artweb.com' virtual art gallery interface [1], which is close to the recommended number of 20 used in a previous study by Nielsen and Sano [23]. These icons were taken at random either individually or from grouped toolbars from various parts of the interface. The icons were designed for various basic interface functions (i.e. carrying out navigation, initiating an action and obtaining information) and are depicted in Table 2, labelled according to their function or purpose.

Table 2: The 21 Icons Evaluated in the Tests

Images of 21 evaluation icons							
1		2		3		4	
	Action		Action		Information		Navigation
5		6		7		8	
	Information		Action		Action		Navigation
9		10		11		12	
	Navigation		Action		Action		Action
13		14		15		16	
	Action		Navigation		Navigation		Information
17		18		19		20	
	Information		Action		Navigation		Navigation
21							
	Navigation						

3.2. Test Sample

Five users consented to take part in the pilot study [1] to evaluate icons by participating in the IIT of icons displayed 'out of context'. The choice of a small sample size in this type of research was based on the studies of icon usability by Nielsen and Sano [23] to collect rich data. The pilot sample included one female and four males - a ratio that is proportionate to the gender balance of the organization in which the tests were conducted. All the participants fell within the age range 20 - 29 years and all had good eyesight and no obvious disabilities.

None of the participants had previously used the 'Artweb.com' virtual art gallery, although 80% had experience of using another virtual tour and had used other 3-D virtual worlds. All the participants had more than ten years' experience of using personal computers and most of the participants fell within the range of 10 to 14 years' experience, as shown in Table 3. This may be because most of the participants in the study were university students undertaking a technology-related degree course.

Table 3: Experience of Using Computers in Years

Years' experience of using a personal computer	
Range of experience	***No. of users***
0 – 4 years	0
5 – 9 years	0
10 -14 years	4
15 -19 years	1
20 -25 years	0

Most of the subjects fell into the range of 30 – 44 hours of weekly computer use, with one subject exceeding 60 hours, as shown in Table 4.

Table 4: Computer Use per Week in Hours

Hours of computer use per week	
Range of No. of hours	***No. of users***
65	0
15 – 29	0
30 – 44	4
45 – 59	0
60 +	1

3.3. Test Procedure

The IIT in the pilot study used a variety of the commonly-used 'card sorting' technique [29]. The participants were provided with brief details of the test scenario as in previous studies of this type [30]. The test administrator then conducted the IIT with the participants individually, each session lasting approximately forty-five minutes [1]. This procedure was repeated and the participant's interpretation of the icons' meaning or function was noted until all 21 cards had been displayed. An overall results table was produced by calculating the IRR expressed as a percentage for each of the icons using the formula:

(No. of correct responses / No. of participants) x 100 = Icon Recognition Rate %.

3.4. Results for Icons 'Out of Context'

The IIT results for all 21 icons tested 'out of context' were placed into the chosen icon classification (i.e. 'identifiable', 'mediocre' and 'vague') based on the participants correctly interpreting their meanings or functions. In the pilot test, fifteen icons (i.e. 71.4% of the set of 21 icons) were classed as 'identifiable', one was classed as 'mediocre' (i.e. 4.8% of the set) and five were classed as 'vague' (i.e. 23.8%). This high proportion of 'identifiable' icons could suggest that the designs were generally successful in this interface. However, the meaning of 28.6% of the icons (i.e. the 'mediocre' and 'vague' classes) was misinterpreted or confused, which could seriously compromise

the usability of the interface in practice. For the purposes of the pilot study [1] a 'traffic light' system was used to indicate the icons' classification according to their IRR score, from best to worst, (i.e. green applies to 'identifiable' icons, amber to 'mediocre' icons and red to 'vague' icons) as in Table 5.

Table 5: Classification of Icons by IRR Score

Classification of icons as identifiable, mediocre or vague					
No.	*Image*	*Meaning*	*Score*	*%*	*Class*
1		Start Virtual Tour	5/5	100.0	**Ident.**
2		Previous tour position, pause tour, next position.	5/5	100.0	**Ident.**
3		Exhibition information	5/5	100.0	**Ident.**
8		Previous artwork to the left	5/5	100.0	**Ident.**
10		Play animation button to circle artwork	5/5	100.0	**Ident.**
11		Pause animation button to circle artwork.	5/5	100.0	**Ident.**
13		Pan and zoom image.	5/5	100.0	**Ident.**
16		Information on artwork.	5/5	100.0	**Ident.**
17		Contact the exhibitor (by email).	5/5	100.0	**Ident.**
19		Navigation arrow buttons	5/5	100.0	**Ident.**
5		Help information for navigation.	4/5	80.0	**Ident.**
14		Next artwork to the right	4/5	80.0	**Ident.**
6		Full screen of virtual exhibition.	3/5	60.0	**Ident.**
7		Return to screen to window size.	3/5	60.0	**Ident.**
18		Close window button.	3/5	60.0	**Ident.**
12		Slider to zoom in & out of image.	2/5	40.0	**Med.**
9		Rotate left (anti-clockwise)	1/5	20.0	**Vague**
15		Rotate right (clockwise).	1/5	20.0	**Vague**
20		Fast jump to location.	1/5	20.0	**Vague**
21		Jump to next room.	1/5	20.0	**Vague**
4		Back to start point of virtual art exhibition.	0/5	0.0	**Vague**

3.5. Findings from the Pilot Study

The pilot study [1] showed that 'universal' icons from applications with which participants were already familiar were easily recognized. Icons that resembled those used in other interfaces and packages, but which had different functions, were confusing to the respondents and did not match their expectations. It was concluded from the pilot study [1] that icons that closely resemble their intended function and therefore do not require prior learning or experience achieve a higher IRR score. The pilot study also showed that icons taken out of context or which have been encountered previously in another context can be confusing to the user. This appears to depend on the user's experience, knowledge and familiarity with that type of interface.

Some icons in the interface *appeared* to be common to most applications (e.g. the 'question mark' suggests a general help function) but were used in this case for an unusual purpose (i.e. specific navigation help) contrary to the user's expectations. Therefore, adding more visual detail to the icons to make them more concrete [19] may help users by reducing their ambiguity. However, it may take longer initially for the users to process the icon's meaning cognitively [16]. In fact, the pilot study suggests that designers' adaptation of the same icon for different purposes appears to be creating misinterpretation. There are also other factors which may influence icon recognition, including the icons' grouping in tool bars, their location on the screen, their function, distinctiveness, color and boldness.

3.6. Implications of the Pilot Study

The purpose of a pilot study is to provide pointers and guidelines so that further research can be carried out more effectively. The pilot study found that although most of the icons tested (15/21 or 71.4%) are 'identifiable', a significant proportion of them are not functioning effectively (see Table 5). Of the icons tested 'out of context' 28.6% (6/21) failed to meet the adopted level of identifiability, which is lower than the ISO standard for signs in general. Of these 'unsuccessful' icons, one was classed as 'mediocre' (scoring 40% IRR) and 23.8% of the total (5/21) were in the lowest 'vague' class, having an IRR of 20% or lower. The meaning of one icon was not recognized by any of the participants (scoring 0% IRR). If these findings are extended to virtual interfaces in general, this lack of recognition could have serious consequences for the effectiveness of icon-driven virtual interfaces in terms of usability, the quality of the users' experience and their satisfaction. It was therefore decided to explore the possibility of extending the research.

Reflection by the researchers and feedback from reviewers offered the following insights into ways in which the pilot study could be extended:

- The small sample size (five participants) inhibited the data analysis. A larger test sample would improve the statistical validity of the recognition test and make it more representative of the real users of a virtual interface. However, the larger sample could make it more difficult to capture the same 'richness' in the data. Nielsen and Sano [23], who devised the tests, justify the use of a sample of five for this reason. In fact, the small sample size means that some values were so marginal that one correct or incorrect interpretation of the icon could increase or decrease the IRR by as much as 20%.
- All the participants were expert computer users, and all had used virtual tour software. This may not be representative of the typical users of a virtual gallery. Similarly, the age range of the participants could be expanded to be more representative of such users. In the pilot study all the participants were in the 20 to 29 age group. A similar study by Gatsou *et al*. [15] that included participants from 20 to 79

suggests that icon recognition declines consistently with age. It would be interesting to test this. The extended research using the same icons should therefore include novice users and older users, which would provide an interesting comparison of the way in which experts and novices and different age groups interpret icon types.

- The test 'out of context' was felt to be a fair assessment of the ability of an icon to convey its meaning, but also to be unrealistic as a test of its success 'in action'. Further tests should therefore be carried out to assess the users' understanding of the meaning and purpose of the same icons when placed in context, which was felt to be a more realistic evaluation of their function in an interface through environmental clues and positioning. The extended research therefore includes more detailed tests of icons and records more data about the ways in which users understand and interpret icons both in and out of context.
- Little was recorded in the pilot study [1] about the factors which may affect individual participants' performance in the test. The findings suggest that a user's personal profile, including factors such as prior knowledge and experience and cognition and learning style, can affect the usability of the interface as well as the degree of 'immersion'. The extended study therefore includes some of these factors and examines them as influences on icon recognition success.

4. The Extended Study

The testing method used in the pilot study [1] was developed from Icon Intuitiveness Testing by Nielsen and Sano [23]. The study indicated that an IIT is a useful tool for assessing how accurately an icon expresses its intended meaning. However, it was felt that the extended study should provide richer data through which the icons could be evaluated in more depth. Experience of the IIT in practice suggested that improvements could be made. The testing method used therefore draws to some extent on all the other methods explored in the pilot study [1] but is adapted for the extended study. The chosen testing method is therefore termed Icon Recognition Testing (IRT) to avoid confusing it with other testing methods.

4.1. Choice of Subject

It was decided that the extended IRT required as a subject an advanced virtual interface with icons having the following features:

- The icons should be capable of different interpretations in and out of context and be used to carry out a range of functions. Ideally these should include 3D navigation and 'jumping' from one location to another, obtaining information about the interface and exhibits and performing action functions such 'zooming' and rotation. They should also initiate sophisticated user-driven interface functions such as screen and window manipulation.
- The icons should be capable of being tested 'out of context' and 'in context' by using small icon cards and still 'screen shots' from the virtual art gallery interface. It was not intended that a fully functional interface should be used, as this may suggest the function of the icons too readily to the participants in the study.
- Some of the icons should be grouped in tool bars as well as being displayed individually and some should only appear when they are usable (i.e. 'toggled').
- The icons are used for the basic activities that a visitor would carry out in a 'real' art gallery (e.g. navigation around the exhibits and obtaining information about the gallery and artworks) as well as virtual 'action' functions (e.g. closing a 'pop-up' menu).

After a selection process failed to identify a superior candidate site, it was decided to use the desktop version of the same German 3-D virtual art gallery (i.e. 'Artweb.com') that had been used for the pilot study. The website is a more 'traditional' type of virtual gallery, using a selection of different styles of room layout based on 'real' art gallery architectural plans. It uses an interactive virtual environment, in which users can navigate through a 3-D space using a mouse and keyboard to access an array of icons to carry out tasks using buttons, cursor pointers and interface metaphors.

This website may be less immersive than some that use high-end interactive technology (e.g. VR headsets, helmets and gloves) but it includes a larger selection of icon types and functions [31]. This makes it more useful for an icon recognition test than some of the later generation of virtual tour interfaces that rely on techniques such as 'swiping' for some of their navigation actions.

It is important to state that the extended study is not a critical test of this specific site, but a general test of the extent to which certain icons convey meaning and of the usability of this generation of virtual gallery sites of which it is typical. The rationale behind the IRT was to gain an insight into how participants from different backgrounds with varying levels of experience and alternative perspectives would perceive the meaning of the icons. Also, it was intended to see if there is a difference in IRR score between the icons seen 'out of context' and 'in context'. In this study, an icon is taken to be 'in context' if two factors apply:

1. There are visual cues in the virtual environment to aid the user in understanding the meaning and/or function of the icons including landmarks, points of reference (e.g. non-interactive objects), contours and boundaries (e.g. walls and doorways), routes around landmarks (e.g. pathways) and room layouts of exhibits [32].
2. Control tool bars are used, with a hierarchical structure, having icons grouped according to their purpose, which change according to the virtual 'position' of the user in the interface or the function being requested.

Although the tests identified in the literature review examined similar aspects of icon understandability [15,21,12] as far as the researchers can ascertain no test has examined the same properties of icon design using the same measures of icon recognition. This extended study is therefore an original contribution to the field of icon design as well as to the construction of virtual interfaces. One implicit purpose of the study is to understand how misconceptions arise and to derive recommendations or guidelines for a more effective way of designing icons, allowing virtual interfaces to be developed that enhance ease of use and improve the quality of the user's experience.

4.2. Test methodology

The complete IRT used in the extended study consists of two recognition tests and two questionnaires, one administered before the tests and one after both tests had taken place, as follows:

- **A pre-test questionnaire**, which contained 13 basic questions to record the participants' demographic data and level of experience of computing in general and virtual interfaces specifically.
- **Test One ('out of context'),** in which participants were shown a range of icons from the interface without any visual cues to their function and were asked to interpret the meaning of each icon. They answered in their own words and their responses were recorded in an Icon Recognition Booklet as brief notes by the Test Administrator.
- **Test Two ('in context'),** in which the participants revisited the icons but were shown the context of the art gallery and the environment in which the icon would be seen. As with Test One the responses were recorded in the Icon Recognition Booklet. The responses to Tests One and Two were then analyzed for themes and are reported as Thematic Analysis 1 & 2.
- **A post-test questionnaire,** which contained a series of 'yes/no' questions in two sections:

Section 1 ('out of context') relating to Test One

Question 1. *Were any of the 21 icons easier to recognize when out of context?*

Question 2. *Were any of the 21 icons harder to recognize when out of context?*

Section 2 ('in context') relating to Test Two

Question 1. *Did viewing any of the 21 icons in context (Test Two) change their meaning from Test One (i.e. out of context)?*

Question 2. *Are you familiar with any of the 21 icons in other contexts (i.e. software and applications)?*

Question 3. *Does grouping icons into tool bars make their meaning clearer?*

The verbal responses to both sets of questions were recorded *verbatim* in brief form by the Test Administrator in the Icon Recognition Booklet. The two tests lasted around forty-five minutes to one hour with each participant and the initial tests were completed within a one-week period, followed by a further round of tests with a different sample following comments from a reviewer. Six participants without postgraduate qualifications who were employed in non-computer related work were tested and their results replaced six postgraduate expert participants. The test environment in all cases was a quiet room with adequate lighting, free from distractions. A description of the IRT procedure was read out from a Briefing Instruction Sheet and participants were informed about the test scenario as in previous studies [20], before being asked to complete the consent form and pre-test questionnaire.

4.3. Icon classification

Three categories of icon were identified according to their intended function, such as; initiating action (e.g. zooming in and out, opening and closing a window), obtaining information (e.g. about an exhibit or the gallery itself) and navigating around the gallery (e.g. moving to the left and right, going forward and back). The set of icons contained some 'familiar' icons, which resembled those used in other interfaces, as well as some more 'obscure' icons, which would be less familiar to the participants. This combination would test whether experienced users could employ existing conventions to aid their recognition and whether misconceptions could arise because of their existing knowledge and familiarity.

4.4. Pre-test questionnaire – participant demographics

All 21 participants in the tests declared themselves to have good eyesight for computer work and all were competent English speakers, although they had different cultures and nationalities. All were regular users of computers for a variety of purposes. The balance of age, gender, education level and employment (including a category for students) in the opinion of the researchers made the sample representative of the probable range of users of a typical virtual art gallery interface. The responses to the demographic questions are described in the following section:

Questions 1 & 2. *What is your age group? What is your gender?*

The age of the participants was noted, as previous research suggests that the ability to recognize icons declines with age [15] and this was to be tested again. For ethical reasons minors (defined as persons under eighteen) were omitted from the study but apart from that the age range and proportions (from 18 to 69 years) broadly reflect that of visitors to UK galleries in 2016 - 2017 [33]. The gender balance was approximately equal (i.e. 10 males and 11 females) which is also representative of the UK population, as shown in Table 6.

Table 6: Ages and gender of the participant sample

Participant sample by age range and gender			
Age range	***No. of users***	***Male***	***Female***
18 - 25 years	6	1	5
26 - 33 years	6	5	1
34 - 41 years	3	2	1
42 - 49 years	2	2	0
50 - 59 years	2	0	2
60 - 69 years	2	0	2
Totals	**21**	**10**	**11**

Question 3. *What is the highest academic qualification you have obtained?*

The participants were asked to declare their highest level of academic qualification (i.e. school certificate, college diploma, bachelor's degree, master's degree or doctoral degree) as it was felt that this may have some bearing on their ability to interpret the meaning of the icons. This is depicted as a 'pie chart' in Figure 2 with the proportion of participants' highest level of academic qualification expressed as a number (in brackets) and a percentage. One participant (4.8% of the sample) had only a school level

qualification, 38.1% had a college Diploma, 28.6% a Bachelors' degree, 23.8% a Masters' degree and 4.8% a Doctoral degree. It is assumed for the purposes of this research that a sample of adults visiting a virtual art gallery will have a similar educational profile.

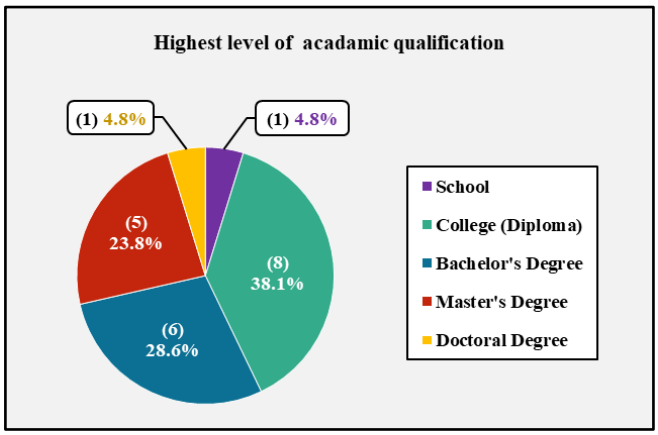


Figure 2: Highest levels of academic qualification

Question 4. *If you are a current or a past student, please state your course title and main area of study.*

The participants' relevant areas of study (e.g. Computing or Art and Design) were recorded briefly in 'free-form' and were placed into seven categories (as shown in Figure 3) to check whether the subjects studied may have some effect on icon recognition. The largest proportion of participants (28.6%) was in the Computing category, with Art and Design the second largest (19.0%), and Sciences constituting the smallest proportion with 4.8%.

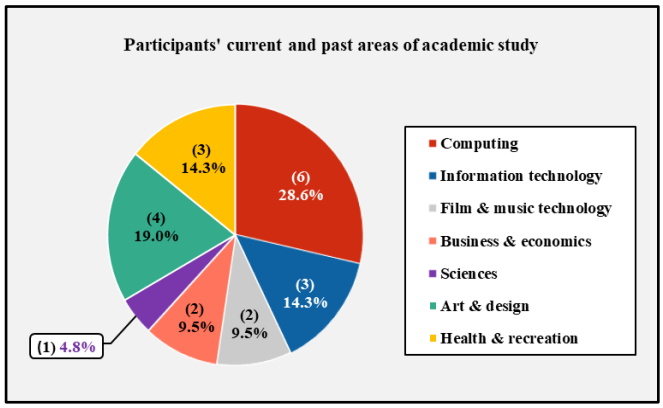


Figure 3: Current and past areas of academic study

Questions 5 & 6. *Which category best describes your occupation? If you are employed, please state your job title.*

Each participant's occupational status (i.e. employed, student, retired or home maker) was recorded with the job category where relevant, to find out if there was a correlation between the participant's employment and his or her ability to interpret icons. It was suspected that certain occupations could develop traits that could affect icon recognition. The primary pie chart on the left of Figure 4 shows the participants' occupational status expressed as a proportion, number and percentage. It is not known whether this employment profile represents the visitors to an actual virtual gallery, but it represents a cross-section of the population.

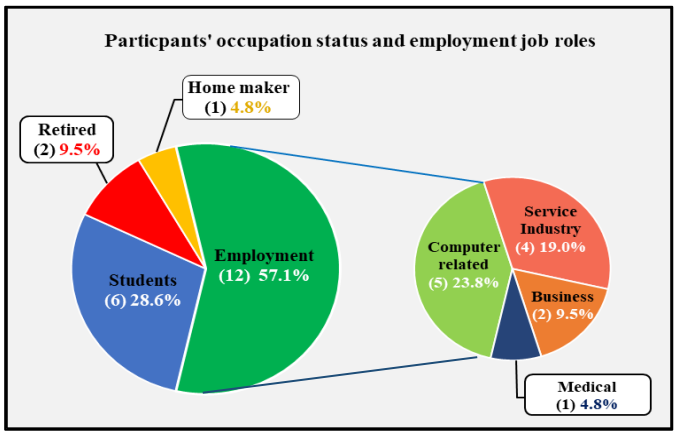


Figure 4: The participants' occupational status

The largest proportion (57.1%) was in employment, while less than a third (28.6%) were students, two people (9.5%) were retired and one person (4.8%) was a home maker. The 'employed' segment was then expanded into a secondary pie chart on the right of Figure 4, which was further divided into job categories, again expressed as a number and percentage. The most common employments were related to the use of computers and the service industries. This implies that less than a quarter of the sample would be regular computer users through their work.

Question 7. *Have you ever worked as an icon designer or a webmaster?*

It was assumed that either of these roles would provide the job-holder with a distinct advantage in terms of icon recognition both 'out of context' and 'in context'. As the sample used in the IRT was intended to be representative of typical virtual gallery visitors, it may be expected that they would have experience as users, rather than as icon creators or designers, so as not to bias the results. It was found that 9.5% of the participants had this type of experience, which was not felt to be excessive. The analysis would show whether experience of icon or website design improved the respondents' ability to recognize the icons.

4.5. Participants' computer experience

Question 8. *Typically, how often do you use a computer interface with icons and for what purpose?*

It was felt that regular use of icon-driven interfaces may have a bearing on the IRR score, so participants were asked to indicate how frequently they used icon-based interfaces and the purpose for which the computer was used, as shown in Figure 5. All the participants used a computer interface daily for Leisure, Home, Work and Study, which constituted the most frequent purpose (i.e. 61.9% of participants). As most packages (e.g. MS Office ®) are icon-driven, this suggests that all the participants would be competent at icon recognition. It should be noted that the operating systems of many commonly-used mobile devices also use an icon-based interface, including Android® and iPhone® mobile 'phones. The retired participants (9.5% of the sample) characteristically did not use computers at all for Work or Study, but used them for Home and Leisure.

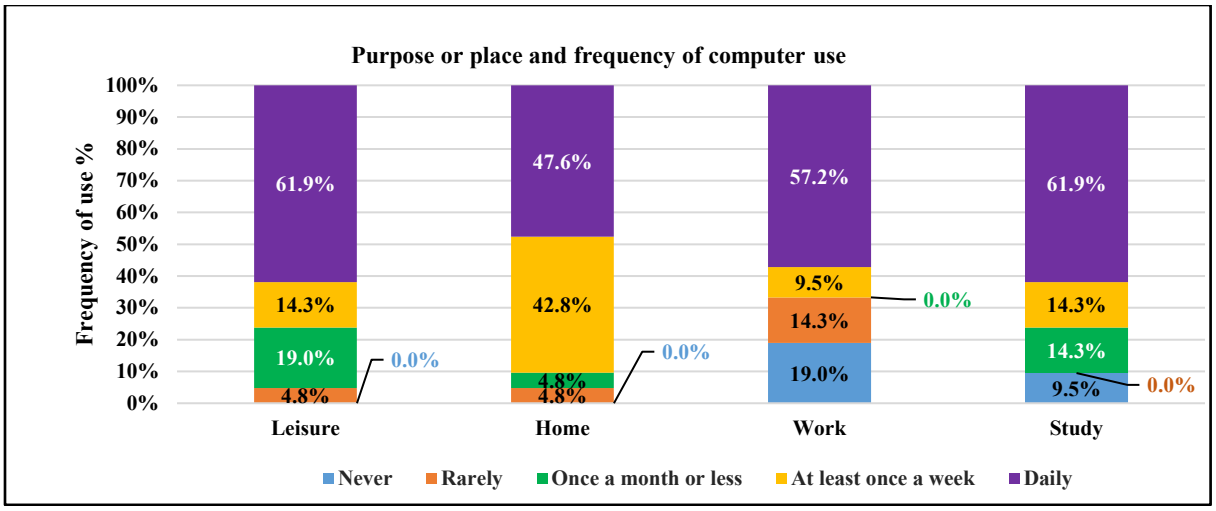


Figure 5: Frequency and purpose of computer use

The responses were given scores of 4 points for daily use, 3 points for use at least once a week, 2 points for at least once a month, 1 point for rarely used and '0' for never used. The point scores for each of the four 'purposes of use' categories were accumulated and the totals were ranked in descending order, as shown in Table 7.

Table 7: Users ranked by frequency and purpose of computer use

Users ranked in order by frequency of computer use					
Users	*Purpose for which computer used*				
	Leisure	Home	Work	Study	TOTAL (pts)
U.3	4	4	4	4	16
U.6	4	4	4	4	16
U.7	4	4	4	4	16
U.14	4	4	4	4	16
U.15	4	4	4	4	16
U.16	4	4	4	4	16
U.21	4	4	4	4	16
U.2	4	4	3	4	15
U.11	4	4	4	3	15
U.5	4	4	4	2	14
U.10	4	2	4	4	14
U.1	2	3	4	4	13
U.19	2	3	4	4	13
U.12	4	3	1	4	12
U.8	3	3	1	4	11
U.20	2	3	3	3	11
U.4	4	3	0	2	9
U.9	3	3	0	3	9
U.17	2	3	1	2	8
U.13	3	3	0	0	6
U.18	1	1	0	0	2
TOTAL	70	70	57	67	264

The highest total score for all categories was the maximum of 16 points (colored green) for Users 3, 6, 7, 14, 15, 16 and 21. The lowest total score was User 18 with 2 points out of a maximum of 16 points (colored red). The median score was 14 and Users 5 and 10 fell into this range, as highlighted by the bold lines. The joint highest frequency of use of computer interfaces was for Leisure and Home (i.e. scoring 70 points) followed closely by Study (scoring 67 points) while Work scored 57 points. The participants overall scored a total of 264 points (78.6%) out of a possible total of 336 points (100%). This indicates that, depending on the types of applications, programs and browsers used, in general the users tested had a significant exposure to a range of icons.

Question 9. *How would you describe your level of computer skills?*

The participants were asked to rate qualitatively their own level of computer skill (rather than their quantitative experience of using computers) as the user's general experience with computers may not necessarily equate with his or her competence in using a virtual interface. The self-described level of computer skill showed that all of the participants had some experience of using computers, 42.9% of the sample describing themselves as 'advanced' and equal percentages (28.6%) rating themselves as 'intermediate' and 'basic' as shown in Figure 6. This can be said to represent a typical range of the computer expertise that would be found in visitors to a virtual gallery.

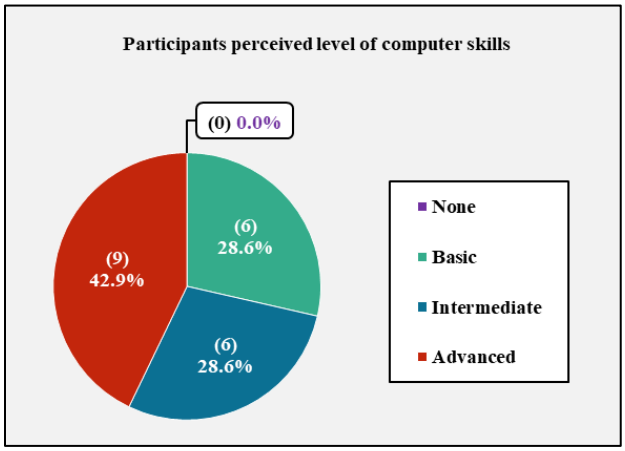


Figure 6: Self-described levels of computer skills

Question 10. *Which of the following devices do you use to access the internet?*

The participants were asked to indicate which of the ten most common computing devices they used to access the Internet, with the opportunity to record less common devices in free-form as 'other'. The responses were given scores of '1' for a ticked box or '0' for an unticked box. The scores for the number of devices for each of the 21 users were added and this total score was ranked in descending order as shown in Table 8.

Table 8: Ranked order of participants by the number of hardware devices used

Users ranked in order by number of hardware devices used												
Users	*Hardware devices*											
	Smart Watch	Smart Phone	PDA	E-reader	Tablet	Notebook	Laptop	Desktop	Games Console	Smart TV	Other	TOTAL
U. 6	0	1	1	1	1	1	1	1	1	1	0	9
U.2	0	1	0	0	1	0	1	1	1	1	0	6
U.10	0	1	0	0	1	0	1	0	1	1	0	5
U.11	0	1	0	0	1	0	1	1	1	0	0	5
U.1	0	1	0	0	0	0	1	1	0	1	0	4
U.14	0	1	0	1	1	0	1	0	0	0	0	4
U.16	0	1	0	0	1	0	1	1	0	0	0	4
U.3	0	0	0	0	1	0	1	1	0	0	0	3
U.5	0	1	0	0	0	0	1	1	0	0	0	3
U.7	0	1	0	0	0	0	1	1	0	0	0	3
U.8	0	1	0	0	1	0	1	0	0	0	0	3
U.9	0	1	0	0	1	0	1	0	0	0	0	3
U.12	0	1	0	0	1	0	1	0	0	0	0	3
U.15	0	1	0	0	1	0	1	0	0	0	0	3
U.17	0	1	0	0	0	0	1	0	0	1	0	3
U.20	0	0	0	0	0	0	1	1	0	1	0	3
U.21	0	1	0	0	1	0	0	1	0	0	0	3
U.4	0	0	0	1	0	0	1	0	0	0	0	2
U.13	0	0	0	0	0	0	1	0	0	1	0	2
U.18	0	0	0	0	0	0	1	0	0	0	0	1
U.19	0	1	0	0	0	0	0	0	0	0	0	1
TOTAL	0	16	1	3	12	1	19	10	4	7	0	73

The participant with the highest score was User 6 with 9/10 devices (colored green) and the joint lowest were Users 18 and 19 with 1/10 devices (colored red). The median score was three devices and Users 3, 5, 7, 8, 9, 12, 15, 17, 20 and 21 fell into this range, as highlighted by the bold lines. In terms of the devices, more participants used laptops (19 users) followed by smartphones (16 users) and tablets (12 users). No-one used the smartwatch and older devices such as PDAs also achieved low numbers (one user). It was noted that 'smartphone' interfaces tend to be icon-driven which could affect the results of the IRT. The desktop version of the virtual interface was chosen for the test as it was felt by the researchers that this version was most likely to be used for virtual tours of a gallery or museum due to the size and quality of the monitor. It is unlikely that artwork would be viewed in detail on a smartphone or even a tablet by discerning art lovers.

Question 11: *Which of the following types of computer application have you used and how frequently?*

The respondents were asked to indicate how frequently they used nine types of computer application, (i.e. regularly, occasionally or never used) to establish their familiarity with different types of interface and their experience of viewing icons in different contexts. The responses were given scores of two points for 'regular use', one point for 'occasional use' and zero for 'never used'. The point scores for each of the nine categories were added and this total was ranked in descending order, as in Table 9.

Table 9: Ranked order of participants by the number of computer applications used

Users ranked in order by number of computer applications used										
Users	*Computer Applications*									
	Interactive websites	Virtual Worlds	Virtual Tours	Social Media	Navigation	Web Browsers	Media Players	Office Apps	Gaming	TOTAL (pts)
U.20	0	1	2	2	2	2	2	2	2	16
U.5	2	1	1	2	1	2	2	2	2	15
U.6	2	2	1	1	1	2	2	2	2	15
U. 2	2	0	1	1	2	2	2	2	2	14
U. 3	2	1	1	1	2	2	2	2	1	14
U.11	1	1	1	2	1	2	2	2	2	14
U. 1	1	0	1	2	2	2	2	2	1	13
U.19	1	1	1	2	1	2	2	2	1	13
U.21	2	0	1	2	1	2	2	2	1	13
U.7	2	0	0	2	1	2	2	2	1	12
U.10	0	0	1	2	2	2	2	2	1	12
U.15	2	0	0	2	2	2	2	2	0	12
U.13	1	0	1	2	1	2	2	0	2	11
U.14	2	0	0	1	1	2	1	2	2	11
U.16	2	0	1	2	1	2	1	2	0	11
U.17	1	0	0	2	2	2	2	1	1	11
U.8	1	0	0	2	1	2	2	1	1	10
U.9	2	0	0	2	2	2	2	0	0	10
U.12	2	0	0	2	1	2	2	1	0	10
U.4	2	0	1	0	0	2	0	0	0	5
U.18	1	0	0	1	0	1	1	0	0	4
TOTAL	31	7	14	35	27	41	37	31	23	246

The highest total score was User 20 with 16 out of a maximum 18 points (colored green) and the lowest score was User 18 with four points out of 18 (colored red). The median score was 12 points and Users 7, 10 and 15 fell into this range, as highlighted by the bold lines. Most users used Web Browsers (scoring 41 points) and Media Player frequently (scoring 37 points) while Virtual Worlds (scoring 7 points) and Virtual Tours (scoring 14 points) were used less frequently. This suggests that the participants would approach the IRT as average users of a virtual interface rather than as experts, which had been identified as a drawback to the pilot study [1]. The researchers noted that the subject interface uses a mixture of icons that would be familiar to the user and ones that had been created specially or adapted that would be unfamiliar.

Question 12: *Have you ever been to a public or private art gallery before?*

All participants except one had visited a real art gallery before and therefore it was felt that a sufficient number would be familiar with the layout and setting within which the 'in context' IRT would take place.

Question 13: *Have you ever visited the German 'Artweb.com' virtual online art gallery interface before?*

None of the participants had visited the virtual gallery site before and so all undertook the tests on an equal footing in this respect. Participants were given the real name of the gallery.

5. Experimental procedure

A stated in Section 4.2 the IRT consisted of two parts. In the first part, the icons were evaluated 'out of context'. In other words, they were not associated with the interface and there were no contextual clues to their function or purpose. In the second part of the test, the icons' context was indicated by using still 'screen shots' taken from the virtual tour of the gallery, but the interface was not accessed [34]. This would place an emphasis on understanding the icon in its context and would be a fairer test of the icons' success in communicating its meaning. The experimental procedures for each test are described below:

5.1. Experimental Procedure – Test One

The test used a variant of the 'card sorting' technique [29] using icon cards each measuring 28mm by 28mm, depicting images of the icons. An example of the test set-up is shown in Figure 7.

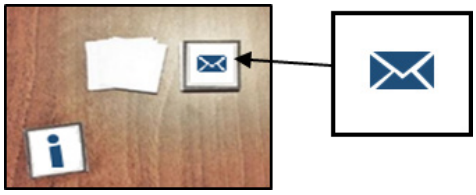


Figure: 7. Setup for the paper-based IRT

In carrying out Test One the following principles were observed:

- The icons included no text [23,25] except for Icon 1;
- The icons were displayed without reference to the actual interface (to preserve the lack of context).
- Only one icon was made visible to the user at a time to avoid giving clues to its use.

The test administrator shuffled the pack of cards to ensure that the icons were not grouped in any way (e.g. by spatial association) before placing them face down on the table as a pack [35]. The administrator then picked up one card at a time from the top of the pile and showed this card to each participant at approximately the same viewing angle and 'reading distance' as it would be in the virtual interface. Each participant was then prompted verbally to attempt a 'free-form' or 'thinking aloud' interpretation of the meaning of each icon [34] as specified in ISO 9186 [26] and following the pattern set by Duarte [36]:

Question 1. *What do you think is the meaning of this icon?*

Question 2. *What function do you think would occur if you clicked on this icon?*

Question *3*. *Does this icon resemble any sign or symbol you have seen or used before?*

The test administrator noted the responses in the appropriate column of the icon recognition booklet *verbatim*. If a participant was not able to interpret the meaning of the icon within one minute, he or she was encouraged to move on to the next icon card and 'don't know' was recorded. It was felt that if users needed this length of time to interpret the meaning of an icon its use in the interface would be compromised. Participants could provide more than one answer and these were noted for later interpretation. After a response was recorded, the test administrator discarded the icon card onto a separate pile, and the participant was not allowed to revisit any of the icons. This process was repeated for all 21 cards.

5.2. Experimental Procedure – Test Two

In Test Two, the same 21 icons were evaluated again but 'in context' (i.e. in their 'natural surroundings'). The participants were shown ten screenshots from the Artweb.com interface on A4 coloured photographic sheets. These screenshots were still images with no interactive functionality and icons were depicted either individually or grouped in toolbars. The participants were therefore able to use visual clues to derive more understanding and meaning from the icons. No text was included, although Icon 1 contained the English word 'Tour'. The ten A4-sized screenshots were shuffled to avoid their functionality being revealed by their sequence or by association. The icons to be identified (singly and in groups) were indicated by red rings [34] as shown in Figure 8.



Figure 8: A screenshot of an art exhibit with tool bars

An Icon Reference sheet showing all the numbered icons was available to the participants and the same testing environment was used as for Test One. Each participant was asked what he or she thought the icon meant (as in Test One) and what purpose the icon had. Participants were encouraged to examine the icon's surroundings for additional clues (i.e. from the gallery room or exhibit) and, where relevant, from other icons that were associated when grouped into tool bars. The test administrator noted the participants' responses in the icon recognition booklet. After a response was recorded, the administrator discarded the screenshot onto a separate pile, face down to avoid influencing the next choice. At the end of the test the participants could use the Icon Reference Sheet to help them fill in the open-ended questionnaire.

6. Scoring criteria for Tests One and Two

After the IRT sessions, the researchers assessed the participants' responses according to the following scoring criteria, adapted from a method developed by Rosenbaum and Bugental [37]:

1. **Completely correct** - the participant's response matches both the object and the function, if not the exact description of the icon's meaning (scored as +2);
2. **Partially correct** - the participant's response matches either the object or the function but not both (scored as +1);
3. **Incorrect** - the participant's response matches neither the object nor the function or the answer is completely different from the intended meaning of the icon (scored as 'zero'). The following cases were included in this category:
   - **a.** Respondent gave 'don't know', 'not sure' or 'no idea';
   - **b.** No response given;
   - **c.** Opposite response given to the true meaning of the icon (e.g. in the case of movement or rotation).

If a participant's entry was not completely clear, a discussion was undertaken by the researchers to interpret the response [35]. In extreme cases the participant was consulted about the meaning. An overall results table giving the IRR score for each icon (shown in Appendix B) was produced by using the following formula, where the maximum possible score for each icon is 42.

(Actual Score / Maximum Possible Score) x 100 = Icon Recognition Rate %.

The IRT results for all 21 icons 'out of context' and 'in context' were separated into classes adapted from a study by Howell and Fuchs [28] with the difference that one class was renamed 'mediocre' instead of 'medium' as it was felt to be a clearer term. The range boundaries differ from those in ISO 3864-2:2016, [38] which refers to general signs rather than computer icons and rates 66.7% and above as 'good'.

According to the Howell and Fuchs stereotypy, icons achieving 60% IRR or above are classed as 'identifiable', whereas icons scoring less than 60% IRR are felt to be 'unsuccessful' in conveying their meaning. The adaptation of this technique that was developed for this research further divides these 'unsuccessful' icons into 'mediocre' (scoring 30% - 59% IRR) and 'vague' (scoring 0% - 29% IRR) as shown in Table 10.

Table 10: Icon Recognition Rate classification

Icon recognition rate (IRR) classification	
IRR (%)	***Classification***
60 – 100	Identifiable
30 – 59	Mediocre
0 – 29	Vague

7. Results of Test One - 'out of context'

The 21 icons used in Test One 'out of context' were given an IRR score according to the procedure described above and were classed as 'identifiable, (60% - 100%), 'mediocre' (30% - 59%) or 'vague' (0% - 29%) according to the adapted classification system. Where 'identifiable' icons also reached the ISO 3864-2:2016 [38] standard of 66.7% for signs, this was also noted in the results table for interest but was not included in the formal classifications.

The textual comments made by the participants were examined to see if they expressed confidence in their interpretation, for instance by giving several alternative answers or by indicating uncertainty in the hesitant way they provided their responses. This was felt to be important in the 'out of context' test as the participants had no other clues to guide them, so the form of the icons alone had to indicate their meaning.

7.1. Test One – 'identifiable' icon results (60% - 100% IRR)

In total, the 'out of context' test produced eight 'identifiable' icons (i.e. Icons 1, 2, 8, 10, 11, 13, 14 and 19) which is 38.1% of all the icons evaluated in the IRT, as shown in Table 11. The icons are presented in the table in numerical order with the score out of a maximum total of 42 (i.e. 2 points for an icon that is 'completely successful' in conveying its meaning) in the fourth column and the IRR% in the fifth column.

Table 11: 'Identifiable' Icons 60% - 100% IRR

Identifiable icons scoring an IRR 60% -100% range				
Icon			***IRR for IRT***	
No.	**Image**	**Meaning**	**Score/ Max. score**	**IRR %**
1		Start Virtual Tour	33/42	78.6%
2		Previous tour position, pause tour, next tour position.	28/42	66.7%
8		Previous artwork to the left	25/42	60.0%
10		Play animation button to circle artwork	28/42	66.7%
11		Pause animation button to circle artwork.	30/42	71.4%
13		Magnifying glass - Pan and zoom image	25/42	60.0%
14		Next artwork to the right	25/42	60.0%
19		Navigation arrow buttons	33/42	78.6%

NB: Icons with an IRR meeting the ISO 3864-2:2016 [38] standard of 66.7% and above are underlined.

7.2. Test One – 'mediocre' icon results (30% - 59% IRR)

In total, there were nine 'mediocre' icons (i.e. Icons 3, 5, 6, 7, 12, 16, 17, 18, 20) which is 42.9% of all icons evaluated in the IRT. All the results for the 'mediocre' icons are listed in Table 12.

Table 12: 'Mediocre' Icons 30% - 59% IRR 'out of context'

Mediocre icons scoring an IRR 30%-59% range				
Icon			*IRR for IRT*	
No.	**Image**	**Meaning**	**Score/ Max. score**	**IRR %**
3		Exhibition information	22/42	52.4%
5		Help - with navigation of system or interface	17/42	40.5%
6		Full screen view of the virtual exhibition	23/42	54.8%
7		Return screen to window size (smaller view)	21/42	50.0%
12		Slider to zoom in and out of image.	17/42	40.5%
16		Information on artwork or exhibit	24/42	57.1%
17		Email - contact the exhibitor or gallery.	23/42	54.8%
18		Close window button.	18/42	42.9%
20		Fast jump to go to location	16/42	38.1%

7.3. Test One – 'vague' icon results (0% to 29% IRR)

In total, there were four 'vague' icons (i.e. Icons 4, 9, 15 and 21) which is 19.0% of all the icons evaluated in the IRT. All the results for 'vague' icons are listed in Table 13.

Table 13: 'Vague' icons 0% - 29% IRR

Vague icons scoring an IRR 0% - 29% range				
Icon			*IRR for IRT*	
No.	**Image**	**Meaning**	**Score/ Max. score**	**IRR %**
4		Back to start point of virtual art exhibition	0/42	0.0%
9		Rotate left (anti-clockwise)	7/42	16.7%
15		Rotate right (clockwise).	9/42	21.4%
21		Jump to next room.	9/42	21.4%

7.4. Summary of IRT 'out of context'

The IRT 'out of context' showed that eight icons of the 21 icons (i.e. 38.1%) achieved an average IRR above 60%. These icons were therefore classed as 'identifiable'. Nine icons (42.9%) scored an average IRR% between 30% and 59% and were classed as 'mediocre'. Four icons (19.0%) failed to reach 30% IRR and were therefore classed as 'vague' (see Table 14). That is not to say that the icons would not function, but it is a strong indication that the user experience would be confusing and less than satisfactory.

Table 14: Summary of icon classes, IRR ranges and results 'out of context'

Summary of icon classes, IRR ranges and results			
Class of icons	**IRR Range**	**Icon Nos.**	**No. & % of icons**
'Identifiable'	60% to 100%	1, 2, 8, 10, 11, 13, 14, 19	8/21 = 38.1%
'Mediocre'	30% to 59%	3. 5, 6, 7, 12, 16, 17, 18, 20	9/21 = 42.9%
'Vague'	0% to 29%	4, 9, 15, 21	4/21 = 19.0%

8. Results of Test Two – 'in context'

All 21 icons shown 'in context' were given an IRR score in the same way as the 'out of context' test and were classified as in Test One. Icons that reached the ISO 3864-2:2016 [38] standard of 66.7% were also noted but not included in the formal classifications. In this case, as with Test One, the verbal responses from the participants were analysed for the degree of confidence they showed in their interpretation of the icons' meaning, for instance by giving several different answers, by the length of time they pondered while providing a response or by the degree of uncertainty they showed in coming to a decision.

This was felt to be important in the 'in context' test as the participants now had clues (e.g. the position of an icon in relation to a landmark or the association of an icon with an exhibit) to guide them. The researchers were interested to see if the inclusion of contextual clues improved the participants' confidence in their decision-making process. However, confidence in reaching a decision about the meaning of an icon is not necessarily associated with the correctness of the interpretation. It is possible to be confident and incorrect. This could apply particularly to icons that are used in a different context from that with which the participants are familiar. This is discussed in the Textual Analysis in Section 9.

8.1. Test Two – 'identifiable' icon results (60% - 100%)

Icons which achieved an IRR score within the 60% - 100% range are classed as 'identifiable'. In total, there are sixteen 'identifiable' icons (i.e. Icons 1, 2, 3, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19 and 20) which is 76.2% of all icons evaluated in the IRT. The 'out of context' IRT had already shown that 29.0% of icons were in the this category. Eight icons have therefore improved their IRR score and moved up to the 'identifiable' category from the 'mediocre'. In the 'in context' test only one of the 'identifiable' icons (i.e. Icon 20) failed to meet the more stringent ISO standard of 66.7% IRR. The results show that twice the number of icons were classed as 'identifiable' in context (16) when compared to out of context (8).

This increase in the participants' ability to recognize the purpose of the icons when the context is known (even in a limited way by showing a screenshot) implies that contextual knowledge makes a significant difference to a users' understanding of an icon's meaning and function. All the results for 'identifiable' icons are listed in Table 15 and Appendix B.

Table 15: 'Identifiable' icons 'in context'

Identifiable icons scoring an IRR 60% -100% range				
Icon			*IRR for IRT*	
No.	**Image**	**Meaning**	**Score / Max. score**	**IRR %**
1		Start Virtual Tour.	35/42	83.3%
2		Previous tour position, pause tour, next tour position.	33/42	78.6%
3		Exhibition information	30/42	71.4%
6		Full screen.	32/42	76.2%
7		Return screen to window size (smaller view).	28/42	66.7%
8		Previous artwork to the left.	35/42	83.3%
10		Play animation button to circle artwork.	31/42	73.8%
11		Pause animation button to circle artwork.	38/42	90.5%
12		Slider to zoom in and out of image.	28/42	66.7%
13		Magnifying glass - Pan and zoom image.	30/42	71.4%
14		Next artwork to the right	36/42	85.7%
16		Information on artwork or exhibit.	33/42	78.6%
17		Email - contact the exhibitor or gallery	29/42	69.0%
18		Close window button.	34/42	81.0%
19		Navigation arrow buttons	38/42	90.5%
20		Fast jump to location.	25/42	60.0%

NB: Icons with an IRR percentage equal to ISO 3864-2:2016 [38] standard of 66.7% and above are underlined.

8.2. Test Two – 'mediocre' icons results (30% to 59% IRR)

Icons which scored an IRR percentage within the 30% to 59% range of the IRT when in context are classed as 'mediocre' and in addition fall below the acceptable level of the ISO standard. In total there are four 'mediocre' icons (i.e. Icons 7, 12, 20, 21) which is 19.0% of all the icons evaluated in the IRT. The results 'in context' show a decrease in the number of icons classed as 'mediocre', as eight icons (i.e. Icons 3, 6, 7, 12, 16, 17, 18 and 20) have now moved into a higher band (i.e. they have become more identifiable when the context is known). One icon (Icon 21) moved into this class from the 'vague' category. None of the icons in this category became less identifiable when the context was made clear. It may be significant that all the 'mediocre' icons appear to have been designed specifically for this interface. Their unfamiliarity therefore gives scope for misidentification and confusion over their meaning and purpose. The IRR scores for 'mediocre' icons are listed in Table 16 and Appendix B.

Table 16: 'Mediocre' icons 'in context'

'Mediocre' icons scoring an IRR 30% -59% range				
Icon			*IRR for IRT*	
No.	**Image**	**Meaning**	**Score / Max. score**	**IRR %**
5		Help - with navigation of system or interface.	21/42	50.0%
21		Jump to next room.	24/42	57.1%

8.3. Test Two – 'vague' icon results (0% to 32%)

Icons which participants scored an IRR percentage within the 0% to 29% range when evaluated in context are classed as 'vague'. In total, there are only three 'vague' icons (i.e. Icon 4, 9 and 15) which is 14.3% of all icons evaluated . The results show that Icon 21 that was vague 'out of context' moved up to the 'mediocre' class 'in context'. The results for 'vague' icons are listed in Table 17 and Appendix B.

Table 17: 'Vague' Icons 'in context'

'Vague' icons scoring an IRR 0% -29% range				
Icon			*IRR for IRT*	
No.	**Image**	**Meaning**	**Score / Max. score**	**IRR %**
4		Back to start point of virtual art exhibition.	5/42	11.9%
9		Rotate **left** (anti-clockwise)	4/42	9.5%
15		Rotate **right** (clockwise).	3/42	7.1%

The IRT 'in context' showed that an awareness of context through seeing the screenshots made a significant difference to the users' ability to recognize the purpose of the icons. In some cases (e.g. Icon 2) this is quite small (a 2.4% increase in IRR) but in most cases, increases in the IRR of between 10% and 20% are achieved. In five cases (Icons 8, 9, 11, 14) the increase in IRR is between 20% and 30% and Icons 18 and 21 both achieved increases of more than 40%. This demonstrates clearly and practically that context plays an important role in icon recognition. The icons within each classification and the proportion of the total icons that they represent 'in context' is shown in Table 18.

The eight icons that were moved into a higher classification through evidence of their context are shown in green in the 'comments' column of the table. Significantly, the icons in the 'vague' category that performed less well in context (Icons 9 and 15) appear to have been designed especially for this virtual interface. Knowing the context in these cases did not seem to help.

Table 18: A summary of icon classes, IRR range, icon numbers and proportions and results 'in context' compared to 'out of context'

Icon classes, IRR range, icon numbers and results 'in context'					
Class	**IRR range**	**Context**	**Icon No.**	**No. & %**	**Comments**
'Identifiable'	60% – 100%	Out	1, 2, 8, 10, 11, 13, 14, 19	8/21 = 38.1%	8 icons (in green) up from the 'mediocre' class
		In	**1, 2, 3, 6, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20**	16 /21 = 76.2%	
'Mediocre'	30% - 59%	Out	3, 5, 6, 7, 12, 16, 17, 18, 20	9/21 = 42.9%	1 icon (in green) up from the 'vague' class
		In	**5, 21**	2/21 = 9.5 %	
'Vague'	0% - 29%	Out	4, 9, 15, 21	4/21 = 19.0%	The 'vague' class lost one icon to 'mediocre' when in context
		In	**4, 9, 15**	3/21 = 14.3 %	

9. Analysis of pre-test questionnaire responses

The respondents' demographic data and their personal profiles (e.g. academic training, experience of interface use, familiarity with computer devices and applications) were recorded in the Pre-test Questionnaire as shown in Section 4.4. An analysis of the data allows interesting comparisons to be made with the results of Tests One and Two. The average of the overall averages (i.e. an average of the 'out of context and 'in context' total IRR scores) is 57.1% as shown in Appendix A.

Questions 1 & 2. *What is your age group? What is your gender?*

An analysis of the responses to Question 1 shows that the findings of Gatsou *et al.* [15], that the ability to recognize icons declines consistently with age, appears to be confirmed. There was one additional observation, that the youngest age group was fourth out of the six, with an overall average of 54.4% (see Table 19) and performed lower than the average of overall averages (i.e. 57.1%) although other factors may have influenced this result. An analysis of the responses to Question 2 shows that, when grouped according to the overall average IRR score, the male respondents performed slightly better than the females (64.0% IRR for males, 50.9% for females). Eight male respondents and three females are above the 57.1% average of overall averages (see Appendix A). It may be implied from this that the males at least in this sample are better at icon recognition than the females.

Table 19: Overall average IRR performance by age range

Overall average IRR performance by age range			
Age range	***No. in group***	***Total overall IRR***	***Overall average.***
18 - 25 years	6	326.4	54.4%
26 - 33 years	6	387.0	64.5%
34 - 41 years	3	182.2	60.7%
42 - 49 years	2	116.7	58.4%
50 - 59 years	2	106.0	53.0%
60 - 69 years	2	81.0	40.5%

Question 3. *What is the highest academic qualification you have obtained?*

It may be assumed that the level of education relates to the user's ability to discern the meaning of icons. An analysis of the responses based on the overall IRR average supports this assumption, but not strongly. Two of the respondents educated to College level scored above the average and six below. Four respondents educated to Bachelors' level scored above the average and two below. Three respondents with Masters' degrees scored above the average and two below. The single respondent educated to Doctoral level scored above the average but not significantly. Therefore, it can be inferred that the user's educational level may have a small influence on icon recognition.

Question 4. *If you are a current or a past student, please state your course title and main area of study?*

It could be assumed that users with qualifications in technical or 'visual' subjects would be better at recognizing icons. Significantly, all five of the top five respondents had qualifications in either Computing, Information Technology or Film and Music Technology, which tends to confirm this. Their skill could be because they had experience of virtual interfaces. The other qualifications were generally distributed among the sample, although it is noticeable that the bottom three scores (well below the average of overall averages) had qualifications in Art and Design and Business and Economics.

Questions 5 & 6. *Which category best describes your occupation? If you are employed, please state your job title.*

It was suspected that experience of certain occupations could affect icon recognition. An analysis of the data shows that 28.6% of the participants were students and 57.1% were employed in various job categories, with two people being retired and one person a home maker. An analysis of the responses to this question indicated that being a student gave only a slight benefit, as the Employed category averaged 58.2% IRR and the Student category averaged 59.0% IRR. The two Retired respondents averaged 40.5% IRR (below the average of overall average IRR of 54.7%) but were by no means the lowest scorers, being in 18th and 19th place. The Home-maker respondent averaged 66.7% IRR and was in sixth place.

Question 7. *Have you ever worked as an icon designer or a webmaster?*

It was suspected that either role would probably have included experience or training that would increase the ability to recognize the meaning of icons. Two participants declared that they had worked in these roles (i.e. 9.5% of the sample). Both achieved IRR scores above the average of 57.1% and were in the top five places. From this we can conclude that experience as an icon designer may improve icon recognition. This may have implications for aligning the designers' conceptual model and the users' mental models when creating virtual interfaces. This mental model is influenced by the user's profile, including his or her experience, interests, learning style and preferences which the designer needs to know.

Question 8. *Typically, how often do you use a computer interface with icons and for what purpose?*

It was felt that regular use of icon-driven interfaces may have a bearing on the IRR score, and some purposes may also favour icon recognition. An analysis of the data in Table 7 shows that seven of the 21 respondents used a computer 'Frequently' (i.e. four points) for all the purposes of Leisure, Home, Work and Study. The data does not show that frequent and varied use is necessarily associated with accurate icon recognition, as three of the seven scored below the average of both overall averages of 57.1% IRR.

Question 9. *How would you describe your level of computer skills?*

It was felt by the researchers that the more skilled in computer use the participants felt themselves to be, the more confident they would be in interpreting the meaning of the icons. An analysis of the data shows that all the respondent felt themselves to have some degree of self-assessed computer skills. However, the difference between the groups was less than may have been expected. The 'Advanced' group achieved an average IRR score (i.e. between 'in context' and 'out of context') of 58.8%, the 'Intermediate' group 55.0% and the 'Basic' group 56.8% (both the latter being below the average). Surprisingly, the Basic group had a slightly higher IRR score than the Intermediate group. So, self-determined computer skills appear to make little difference to icon recognition, as the score for the Advanced group is only slightly above the average for the whole sample.

Question 10. *Which of the following devices do you use to access the internet?*

This question required the participants to indicate how many and which of the ten most common devices they used. It could be assumed that familiarity with more devices increased the user's experience of different interfaces and types of icons, which could increase the IRR score. In fact, the IRR scores shows no significant correlation between the number of different devices used and the user's ability to recognize icons in the two tests. Indeed, the participant with the highest IRR score (User 21 with an average of 76.2%) used only three devices, and seven of the ten highest scoring participants used less than five (i.e. half the available number of devices). It should be noted that the icons in the test were taken from the desktop version of the virtual interface, which makes it different to the small hand-held devices.

Question 11: *Which of the following types of computer application have you used and how frequently?*

By this question the researchers sought to ascertain if the number of different applications used and the frequency of their use had any effect on icon recognition. Nine different types of application were specified, and points were allocated for each and for the frequency of use. It is possible to see a definite correlation between the variety and frequency of use of computer applications and the IRR score. Eight of the top ten highest scorers in terms of IRR percentage had 12 or more points on the scale (see Table 9).

10. Comparison of icons in and out of context

The results of Tests One and Two were examined and discussed among the researchers. Their interpretations of the findings for each icon are included in the following comparative sections. The comments on the 'out of context' and 'in context' results are followed by a textual analysis of the 'free-form' notes taken from the Icon Recognition Booklet (see Section 4.2).

10.1. Icon 1 - Start virtual tour

Out of context, Icon 1 scored 78.6% IRR in the test (see Appendix B) and was therefore classed as 'identifiable'. There were 14 'completely correct' and five 'partially correct' responses, with two 'incorrect' responses. Interestingly, both of these gave a 'don't know' response, which is rather surprising as its purpose (i.e. the word 'tour') is stated on the icon.

In context, the IRR score for Icon 1 increased to 83.3%, raising it even higher in the 'identifiable' category with 15 participants identifying its meaning correctly. There were five 'partially correct' responses and only one 'incorrect' response, which registered a 'don't know' verdict.

A textual analysis of the 'free-form' responses showed that 'out of context' many of the respondents identified the icon correctly (the word 'Tour' was clearly seen) but did not appreciate its true function as *starting* the tour. A circular arrow on the icon caused confusion, with 'slideshow', 'presentation' and even 'headphones' (which are sometimes used on 'real' gallery tours for audio commentary) being offered as possible functions. 'In context' the respondents were able to assign a more accurate meaning to the icon by seeing it in its 'natural surroundings'.

10.2. Icon 2 - Previous, pause & next on tour

Out of context, Icon 2 scored 66.7% IRR in the test (see Appendix B) and is therefore classed as an 'identifiable' icon. Seven participants were 'completely correct', 14 participants were 'partially correct' and there were no 'incorrect' answers.

In context, the IRR score increased to 78.6%, raising it slightly in the 'identifiable' class, with twelve participants being 'completely correct' and nine participants giving a 'partially correct' estimate and no 'incorrect' responses.

A textual analysis of the 'free-form' responses shows that 'out of context' participants assigned a meaning to the icon based on symbols with which they are already familiar - audio and/or video controls. The use of these symbols goes back to the introduction of the cassette tape recorder in 1963 by Phillips NV. They have since become almost universal, so most of the participants have 'grown

up with them'. Knowing the context in the 'in context' test enabled many respondents to provide more detailed, more informed responses, which increased the IRR score for the icon.

10.3. Icon 3 – Virtual exhibition/gallery interface information

Out of context, Icon 3 scored 52.4% IRR in the test (shown in Appendix B) and was therefore classed as 'mediocre'. Only two candidates were 'completely correct', 18 were 'partially correct' and one 'incorrect'.

In context, the icon's IRR score increased significantly to 71.4%, and it has moved well into the 'identifiable' category with nine 'completely correct' and 12 'partially correct' responses, and no 'incorrect' estimates of its meaning.

A textual analysis of the free-form comments showed that 'out of context' the majority of respondents realised that the use of the letter 'i' was for providing information but were unsure about its exact purpose. In this application the 'i' is for general information about the gallery interface, although this is not clear from the use of a grey colour for the icon. One of the participants thought it was a notification symbol, even though this is normally an 'exclamation mark' in other applications. Placing it into context (i.e. on the main toolbar in the screenshots) no doubt allowed the users to deduce that it referred to information about the gallery, rather than to a specific exhibit. This shows the importance of the proximity of an icon to its function (i.e. its literal rather than semantic distance) or its position in relation to other objects.

10.4. Icon 4 - Back to start point

Out of context, Icon 4 was the least recognized icon of all 'out of context', scoring 0% IRR (shown in Appendix B) and is therefore classed as 'vague'. All 21 participants were 'incorrect' and out of those only one gave a 'don't know' response.

In context, the meaning of the icon was slightly more recognizable, with an IRR of 11.9%, however is still classed as 'vague'. A single participant provided a 'completely correct' response, three gave 'partially correct' responses and 17 participants were 'incorrect', including two 'don't know' responses.

A textual analysis of the free-form responses showed that 'out of context' some participants confused it with icons having a different function in other software packages. Most respondents confused the icon with a MS Vista® loading or buffering button. The shading of the icon (it appears lighter on the bottom) may have given the impression of rotation, which is a feature of loading symbols. Some thought it was for 'brightness' or 'no internet connection' or made wild guesses (e.g. 'sunshine'). 'In context', many responses show the same incorrect assumptions, with 'loading' and 'brightness' responses being frequent.

10.5. Icon 5 – 'Help' with navigation of system or interface

Out of context, Icon 5 scored 40.5% IRR (see Appendix B) and is therefore definitely 'mediocre'. Four participants recorded 'completely correct' responses, nine gave 'partially correct' responses and eight were 'incorrect' without a 'no response'.

In context, the icon IRR rose to 50.0% and so was still in the 'mediocre' category. There were still four 'completely correct' responses while the number of 'partially correct' responses had increased to 13, with four 'incorrect' interpretations (without any 'don't know' responses but with one 'no response').

A textual analysis of the free-form comments showed that 'out of context' most participants identified the basic meaning of Icon 5 with the universal symbol for 'Help' as they are already familiar with it in other contexts without recognising its specific meaning in this application. 'In context', that there were now only four incorrect responses indicates that knowing the context had helped some respondents to improve their estimate of its meaning.

10.6. Icon 6 – Full screen

Out of context, the icon achieved a 54.8% IRR (see Appendix B) and is therefore classed as 'mediocre'. Eight participants were 'completely correct' and seven were 'partially correct'. Six responses were 'incorrect' (with no 'don't know' responses) and one participant gave the opposite meaning (i.e. shrink screen).

In context, the IRR for this icon rose to 76.2% making it clearly 'identifiable'. Twelve respondents were 'completely correct' and eight were 'partially correct' in their estimates. The number of 'incorrect' responses was one without any 'don't know' answers, indicating confidence on the part of the respondents.

A textual analysis of the responses 'out of context' showed that many participants thought that the icon was something to do with navigation (due to the use of arrows). Several users thought it was a 'click and drag' or movement control button and one user thought that it resembled an icon used in Google Maps® for a different purpose. Conventional 'screen adjustment' controls often use overlapping large and small rectangles as icons, showing that standards set by the designers of the most popular applications create *de facto* paradigms that users recognize. In context, most respondents recognized that the icon had something to do with expansion or enlargement but did not know the full functionality.

10.7. Icon 7 – Return screen to window size

Out of context, Icon 7 achieved a 50.0% IRR (see Appendix B) and is clearly classed as 'mediocre' (i.e. slightly less than its opposite Icon 6). Seven responses were 'completely correct' and seven were 'partially correct'. Seven were 'incorrect' including one 'don't know', one 'no response' and one 'opposite meaning'.

In context, the IRR for this icon rose markedly to 66.7%, making it 'identifiable' according to the adopted scoring system. Eleven responses were now 'completely correct' and six were 'partially correct' with four 'incorrect' judgements without any 'don't know' responses, indicating confidence if not correctness on the part of the respondents.

A textual analysis of the free-form comments showed that when taken 'out of context' one respondent thought the icon referred to a meeting place or a central point in the virtual gallery. They may have been influenced by the similarity of the sign to the familiar 'assembly point' emergency warning sign (see Figure 9). In context, one of the 'incorrect' respondents thought that the icon meant a return to a point on the virtual tour as the arrows were converging and one thought it enabled the visitor to 'enter the picture'. This suggests that familiarity with common physical signs (in this case the 'assembly point' sign) can create confusion

in the user's mind if icons have a similar appearance but are meant to convey a different meaning.

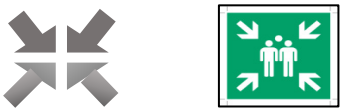


Figure 9: Confusion between 'return to small window' icon and 'emergency assembly point' sign (ISO, 2011)

10.8. Icon 8 - Previous artwork/exhibit to the left

Out of context, the icon scored 60.0% IRR in the test (see Appendix B) and is therefore just classed as 'identifiable'. There were five 'completely correct' and 15 'partially correct' responses. Only one participant identified it incorrectly and in this case the response was the opposite of the intended meaning by recording 'go *forward* to visit next page'.

In context, the IRR rose to 85.7%, one of the largest increases, due to context making it clearly 'identifiable'. There were now 16 'completely correct' estimates and four 'partially correct'. One participant still assigned an incorrect meaning but there were no 'opposite meanings' (zero scores).

A textual analysis of the written responses to Test One showed that the participant who had assigned an opposite meaning to the icon 'out of context' is from a culture which conventionally reads from right to left. This demonstrates that similar virtual interfaces may be intentionally universal in their application, but the icons that control their use are necessarily cultural in their interpretation. In context, the same respondent gave an incorrect (but not opposite) answer, showing that knowing the context suggested a change of interpretation that overcame the cultural expectations.

10.9. Icon 9 - Rotate left (anti-clockwise)

Out of context, Icon 9 scored 16.7% IRR (see Appendix B) and is therefore classed as 'vague'. There were two 'completely correct' and three 'partially correct' responses but these were completely outweighed by 15 'incorrect' responses, of which three had 'opposite' meanings (i.e. rotate in a *clockwise* direction). Interestingly, no-one recorded 'don't know', which shows that the respondents were confident but mistaken in their interpretation.

In context, the IRR dropped to 9.5%, making it even more 'vague' and being the joint lowest score in the test. This low score was created by one 'correct' response, two 'partially correct' responses and 18 'incorrect' responses with seven 'opposite' directions being assumed and three 'don't knows' recorded.

A textual analysis of the free-form comments suggests that many participants identified the icon as initiating a rotation but mistook the direction (perhaps the concepts of 'clockwise' and 'anticlockwise' are less relevant today). Others confused the icon with a 'redo' button (although it was flipped horizontally) or a 'refresh' button, which is like one of the paired arrows from other software packages as shown in Figure 10. In context, several participants assigned meanings that are logically incorrect, such as 'skip forward' and 'go to previous (artwork)' showing their uncertainty. It is also significant that the number of 'incorrect', 'opposite' and 'don't know' responses increased so, knowing the context within which the icons would be used clearly confused some of the respondents. In context, the fact that Icon 9 and its opposite Icon 15 were in toolbars on the opposite sides of the screen to what would be expected (i.e. left-hand rotation on the right tool bar and vice versa) caused the direction of rotation around the artwork to be mistaken.

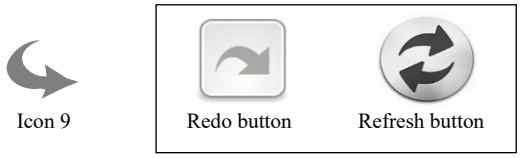


Figure 10: Confusion between Artweb.com 'rotate left' and common 'redo/refresh' icons

10.10. Icon 10 – Play animation button

Out of context, the icon scored 66.7% IRR in the test (see Appendix B) and is therefore classed as an 'identifiable' icon. Nine participants were 'completely correct' in their interpretation, with ten 'partially correct' and two 'incorrect. One of the incorrect responses was because the participant left the answer blank.

In context, the IRR score for this icon (which is 'toggled' with Icon 11) increased slightly to 73.8% and stayed in the upper category, classed as 'identifiable' with 14 'completely correct' and three 'partially correct' estimates, but the number of 'incorrect' responses interestingly increased from two to four.

A textual analysis of the accompanying responses suggested that most of the participants could translate inferences from other sign systems and media objects (e.g. audio or video players) to identify the purpose of the icon. As with Icon 2 (which also originated in the cassette players of the 1960s) an analysis of the free-form comments showed that most of the participants were familiar with its use in domestic audio equipment. In context, the scenario shown was a still image of a sculpture that could be rotated. Two participants gave a new but incorrect meaning to the icon when shown screenshots, confusing the icon's action function with navigation ('go to the right' and 'go to next picture').

10.11. Icon 11 – Pause animation button

Out of context, Icon 11 scored 71.4% IRR in the test (see Appendix B) and is therefore clearly classed as an 'identifiable' icon. There were ten 'completely correct' and ten 'partially correct' responses. Perhaps surprisingly, the one 'incorrect' response had the opposite estimate of its meaning – to start.

In context, the IRR score of Icon 11 (intended to be 'toggled' with icon 10) rose appreciably to 90.5% - one of the highest scores in the test. There were 17 'completely correct' responses and four 'partially correct'. Clearly, placing the icon into context has radically changed the participant's understanding of it.

As with Icon 10 and Icon 2, a textual analysis of the comments following Tests One and Two shows that the participants made similar inferences in evaluating the purpose of the icon. This is another icon that owes its existence to the early tape recorder, representing two tape rollers on a 'reel to reel' tape deck. It is commonly used in domestic sound and video equipment to pause a player temporarily until it is restarted by pressing the play button (i.e. Icon 10). This 'universal' icon's features were unique and is

unlikely to be confused with other icons in the test, although the one user who interpreted it with an 'opposite' meaning thought it was a 'restart' icon. In context, the participants showed a greater understanding of its meaning by associating it with Icon 10 through their familiarity with domestic equipment.

10.12. Icon 12 – 'Slider' to zoom in and out of image

Out of context, Icon 12 scored 40.5% IRR in the first test (see Appendix B) and is therefore clearly classed as 'mediocre'. Seven participants were 'completely correct' in their interpretation of its meaning, three were 'partially correct' and 11 were 'incorrect'.

In context, the IRR for the icon increased noticeably to 66.7% with eleven 'completely correct', six 'partially correct' and four 'incorrect' responses, placing it clearly in the 'identifiable' class.

An analysis of the free-form responses to Test One suggests that this icon is ambiguous, as it was misinterpreted by nine participants as a 'volume control sign' with a slider to change the sound level. Three participants thought it was a 'battery life or power level' indicator rather than a 'zoom slider' due to its similarity to the icon used for this function on some popular devices (see Figure 11). In gaming, it is often used as an indication of a player's energy or power level and in mobile devices it can indicate signal strength, making its use in this context confusing.



Figure 11: Confusion between Artweb.com 'zoom control' and common 'battery power level' icons

In context, Icon 12 appears in the right-hand tool bar when viewing a painting with the associated 'magnifying glass' (Icon 13). This may have clarified its purpose when seen 'in context'.

10.13. Icon 13 – 'Magnifying glass' to pan and zoom image

Out of context, Icon 13 scored 60.0% IRR (see Appendix B) and is therefore narrowly classed as 'identifiable'. The ability of the participants to interpret the meaning of this icon is sharply divided. Nine participants were 'completely correct' and seven were 'partially' correct in their interpretation. On the other hand, there were five 'incorrect' estimates, indicating that this a difficult symbol for some users to recognize decisively when out of context.

In context, the IRR score improved markedly to 71.4%, making it solidly 'identifiable' and showing that knowledge of its context caused the meaning of the icon to be much clearer to most participants. The increase was produced by 15 'completely correct' and six 'partially correct' responses with one 'incorrect'.

A textual analysis of the free-form entries suggests that this icon is confusing, as the 'magnifying glass' part of the sign is 'solid' rather than the ring-like or 'transparent' device used in other common packages (e.g. Photoshop®). Therefore, 'out of context' there were a variety of misconceptions about its meaning. Some of the users saw the icon as a key or a screwdriver symbol, indicating security settings. One saw it as a 'stop sign' and three (no doubt influenced by the common use of a 'magnifying glass' in search engines) saw it as a search function. 'In context', one Muslim participant misinterpreted it as a 'Christian cross symbol' as it was positioned near a painting of the interior of a cathedral. One participant thought it was a 'search/find symbol' but instead of looking for information in a search engine it was looking at specific details in the painting.

10.14. Icon 14 - Next artwork to the right

Out of context, Icon 14 scored 60.0% IRR (see Appendix B) and is therefore narrowly classed as 'identifiable'. There were five 'completely correct' responses and 15 'partially correct'. Participants appeared to have confidence in their judgement, as there were no 'don't know' responses. The only 'incorrect' respondent interpreted the sign as rotation, but in the opposite direction to its intended meaning.

In context, the IRR increased to 85.7% showing that it was clearly 'identifiable' by most participants when seen in its surroundings. There were now 16 'completely correct' and four 'partially correct' interpretations and the respondent who gave the 'opposite' response 'out of context' now had the correct direction.

A textual analysis of the responses showed that 'out of context' the participant who gave the incorrect 'opposite' answer thought the icon's meaning was to 'go back' instead of 'go forward', as the person is from a culture which writes from right to left and made the same mistake with Icon 8, thereby emphasising that cultural interpretations should be taken into consideration when designing interfaces. In context, it was shown that some of the participants did not understand the meaning of the icon fully as they stated the direction as 'go right' instead of the 'next artwork on the right'. One participant who gave an 'incorrect' answer thought the icon's meaning was to focus on the right-hand side of the painting itself.

10.15. Icon 15 – Rotate to the right (clockwise)

Out of context, Icon 15 scored 21.4% IRR (see Appendix B) and is therefore clearly classed as 'vague'. There were four 'completely correct' answers and only one 'partially correct'. An unusually large number of participants (16) gave 'incorrect' answers, out of which four gave 'opposite' meanings (i.e. rotation in an *anticlockwise* direction). Surprisingly, the respondents showed a high degree of confidence in their understanding of the icon as there were no 'don't know' responses.

In context, this icon had an IRR of 7.1%, showing a notable decline. This was caused by the icon receiving only one 'completely correct' response when its context was known. There was again one 'partially correct' interpretation, but there were now 19 'incorrect' responses and the number of 'opposite' directional interpretations (scored as 'incorrect') had increased to seven. Clearly, knowledge of the icon's context had confused the users!

A textual analysis of the free-form interpretations showed that (as with its opposite Icon 9) many participants identified the purpose of the icon but mistook its direction of rotation 'out of context'. Also, there was confusion with an 'undo' button, a 'refresh' button or a 'return' button, which use similar symbols (see Figure 12). In context, the position of the icon in a tool bar on the opposite side of the screen to what might have been expected

probably created confusion as to the direction of its rotation (see also Icon 9).

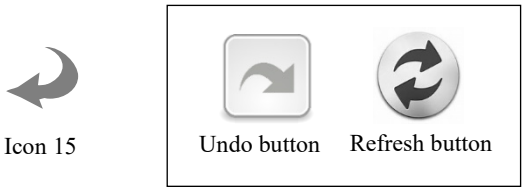


Figure 12: Confusion between Artweb.com 'rotate right' and common 'undo/refresh' icons

10.16. Icon 16 - Information on artwork or exhibit

Out of context, Icon 16 scored 57.1% IRR and was classed as 'mediocre' by a narrow margin. Three participants were 'completely correct' and 18 were 'partially correct'. There were no 'incorrect' responses, indicating only a partial success for this 'universal' icon in communicating its meaning.

In context, like Icon 3 this icon increased its score by a wide margin, achieving 78.6% IRR, making it clearly 'identifiable'. This was largely the result of an increase (to 12) in the number of 'completely correct' responses. There were nine 'partially correct' interpretations and again no 'incorrect' responses.

Textual analysis of the free-form responses showed that 'out of context', all the participants recognized that both Icon 3 (which is grey) and Icon 16 (which is blue) represent 'information' signs but their use tended to elude some of them. This typifies the definition of 'mediocre' icons - the users felt that they knew their meaning but could not work out what to use them for in this interface. 'In context' the IRR score increased considerably, suggesting that association with an artwork helped the participants to identify the purpose of the icon much more accurately. Interface designers should bear this in mind.

10.17. Icon 17 - 'Email' contact the exhibitor or gallery

Out of context, this icon scored 54.8% IRR (see Appendix B) and is therefore classed as 'mediocre', indicating that three participants were 'completely correct' and 17 were 'partially correctly' in their interpretation of its meaning. There was one 'incorrect' response.

In context, Icon 17 increased its IRR score to 69.0%, placing it solidly in the 'identifiable' category. The number of 'completely correct' answers increased to 11 and the number of 'partially correct' responses was now seven. Interestingly, the number of 'incorrect' responses increased to three as context apparently introduced new ambiguity.

Textual analysis of the free-form answers showed that 'in context' most participants identified the basic meaning of Icon 17 with the universal symbol for email. It also shows that some participants could not work out whether the icon was to open an email reader to send or receive an email message and were therefore unable to decide to whom the email was to be sent and about what. This appears to be a case of using a common icon for an unusual purpose, which shows that the design of an icon needs to be aligned with the user's experience and familiarity with similar signs. In context, the email's precise purpose of contacting the exhibitor or gallery about the exhibit became more apparent by its closeness to the exhibit and its association with other icons in the same tool bar that are used to directly to manipulate the exhibit. Although, with experience of its use in social media applications, some users thought it was a way of posting comments.

10.18. Icon 18 - Close window button

Out of context, Icon 18 scored 42.9% IRR (see Appendix B) and is clearly classed as 'mediocre'. Seven participants identified the meaning of the icon completely correctly and four partially correctly. There were ten 'incorrect' responses from participants who attempted to guess the icon's meaning.

In context, the icon's IRR score almost doubled to 81.0%, moving it well up into the 'identifiable' class. This can be attributed to the increase in 'completely correct' responses to 15, while the 'partially correct' responses remained the same at four and two participants registered an 'incorrect' response, one assuming that it marked an observation point and the other a warning.

This is a clear indication that knowledge of context has enabled many participants to improve their understanding of the meaning of the icon. Textual analysis showed that initially 'out of context' there was a wider interpretation by participants with a number of different meanings for a type of warning sign such as a 'no entry', 'stop sign', 'error sign', 'cancel sign' or 'gallery closed' sign. This represented a mismatch with the participants' expectations based on their experience of the symbol in other applications. In fact, the basic form of the icon (although not necessarily its colour) is commonly used to close pop-up windows in a variety of applications, including MS Word® (see Figure 13). An analysis of the free-form responses showed that 'in context' the icon's position on the corner of a window 'frame' made its purpose clearer to the participants, as this is where they would normally expect to see a 'close window' button. This demonstrates the value of consistency, not just in the appearance of icons but in their position and their association with other parts of the interface.



Figure 13: Similarity between 'close pop-up window' icons (left; 'Artweb.com', right; MS Word ®)

10.19. Icon 19 - Navigation arrow button

Out of context, Icon 19 scored one of the highest results with a 78.6% IRR (see Appendix B) and is therefore clearly classed as 'identifiable'. In all, 15 participants interpreted the meaning of the icon completely correctly and three were 'partially correct', while three were 'incorrect'.

In context, the IRR score increased to 90.5%, with 18 participants giving a 'completely correct' response, two 'partially correct' and one 'incorrect', making the icon one of the most identifiable.

Textual analysis of the free-form responses shows that 'in context' some of the participants felt the sign to be like one used in Google Maps® and they therefore interpreted it as a map controller (i.e. for moving a map around a window) rather than a direction control icon (i.e. moving the user's viewpoint). One participant confused this icon with a similar icon often used to enlarge an image or screen. This icon is also familiar to participants

who have experience in playing games which use this type of 3-D navigation tool to move around screens. The context, however, ruled out any association with maps and its position on the screen indicated that it was a navigation tool to most participants.

10.20. Icon 20 - Fast jump to location

Out of context, Icon 20 performed relatively poorly, being placed in the 'mediocre' category with an IRR of 38.1% (see Appendix B). Two participants gave 'completely correct' responses and 12 were 'partially correct' with their estimates. There were seven 'incorrect' answers and out of these, one participant gave a 'don't know' response rather than guessing.

In context, the icon moved up to the 'identifiable' category, although it only just met the criteria with an IRR score of 60%. This was largely because of an increase to eight in the number of 'completely correct' responses. There were nine 'partially correct' estimates and the 'incorrect' responses numbered four. Out of these, the number of 'don't know' answers increased to two and one of these participants had a made a wild guess previously 'out of context', which they then discounted.

A textual analysis of the free-form responses showed that many participants mistook this icon 'out of context' for a 'map pin' marking a specific point rather than a navigation aid, based on its similarity to the marker used in Google Maps® and similar applications with which they were familiar. One interpretation given was as a sign for the start of a tutorial. In context, two participants thought it was a marker for the current location and did not know it was a navigation aid to fast jump to another location in the art gallery. Some of the participants stayed with their originally answers 'out of context', although their responses were more descriptive and related to the context. As the meaning is not clear in or out of context, it seems that the icon requires the user to gain experience with the interface, to learn its functionality.

10.21. Icon 21- Jump to next room

Out of context, Icon 21 scored 21.4% IRR (see Appendix B) and is in the 'vague' category. Only one participant gave a 'completely correct' answer, whilst seven gave 'partially correct' estimates. There were 13 'incorrect' responses and out of those two were 'not sure' or had 'no idea', two gave no response and the rest gave a different meaning or a wild guess as to its purpose (e.g. an architectural feature).

In context, the icon's IRR score rose significantly to 57.1%, moving it up a category to the upper end of the 'mediocre' class. There were now eight 'completely correct' and eight 'partially correct' responses, while five respondents gave 'incorrect' estimates of the icon's meaning.

A textual analysis of the free-form answers showed that some participants mistook the icon for a military insignia 'out of context', as a similar icon appears in many computer games that they had played previously. In context, some participants took it for a sign pointing up to the next floor of the art gallery (e.g. a sign for a lift or elevator), rather than for 'jumping' into the next room on the same level. One participant thought it was an end-point in the gallery visit which could be saved, to allow them to return to the same point. This icon probably requires some prior familiarity through learning the interface to know its functionality.

11. Thematic Analysis 1 - findings 'out of context'.

Questions 1 and 2*. 'Are any of the icons a) easier, b) harder to recognize out of context?'*

The responses to this question suggest that knowledge of context does increase the IRR score but perhaps not as much as previous work [39] would suggest. In some cases, knowing the context made identification more problematic. Icons 9 was felt to be harder to recognize 'in context' by six respondents (28.6% of the sample) while Icon 15 was felt to be more difficult by five respondents (23.8%). These results show that context cannot be relied on to make an icon more understandable, as context can be misleading. With Icons 4, 6, 7, 9, 12, 13, 17 and 18 some respondents who thought that context made identification easier were in fact incorrect in their interpretation. However, knowing the context did enable more accurate recognition in many cases, as was expected. Icons 3, 6, 7, 12, 16, 17, 18 and 20 were felt to be easier to identify through familiarity, and the fact that they moved from the 'mediocre' to the 'identifiable' class bears this out (See Table 19 and Appendix B).

Thematic Analysis 2 - findings 'in context'.

Question 1. *'Do any of the 21 icons change their meaning from what you expected 'in context?'*

The textual analysis showed that being seen 'in context' changed the meaning of all the icons except Icon 8. However, the difference between the 'out of context' and 'in context' tests is lower than expected, in two measures. The first measure is the increase or decrease in IRR between the two tests (as shown in Appendix A). The second is the number of respondents seeing a change in the meaning of the icons when taken 'in context'. Overall, the change is minor and the icon with the largest percentage of respondents is Icon 12 with 47.6% (10/21) believing that context changed the icon's meaning. Only two icons were regarded as having changed their meaning by more than a third of the participants (i.e. Icons 12 and 21). As may be expected, Icon 21 had the largest increase in IRR (+40%).

Question 2*. 'Are you familiar with any of the icons in other contexts?'*

Icon 1 was distinctive and was not confused with other icons. Only one respondent reported seeing something similar on another (un-named) virtual interface. Icon 2 was felt to be like a video, music or media player control by 13 respondents. Interestingly, three thought it was used on YouTube®, but they are not correct. Icon 3 was related to an information function by eight respondents, but none suspected its secondary meaning, which is to give general information about the site. One respondent thought that it was 'greyed out' because it was not active. As the icon cards were displayed in random order, some respondents may have already seen the 'blue' information icon, though none mentioned it. Icon 4 was perceived as a loading/buffering symbol or a brightness control by twelve respondents and as 'settings' (often represented by a 'gear' icon) by four. No-one suspected that it was intended to return the user to the start of the tour. Icon 5 was seen to be a common 'help' icon with confidence by five respondents, and one felt that it served the same function as a in Microsoft® applications. Two respondents offered the extra information that it offered help about the tour.

Icon 6 was felt to be a 'full screen' icon by five respondents and two offered the secondary information that it was like a YouTube® control (not correct). The same respondents felt that Icon 7 was a 'shrink screen' control and again mistakenly attributed it to YouTube®. Icon 8 was likened to the 'go back' icon on websites, etc. by eight respondents. Four identified it correctly as referring to the previous artwork. Icon 9 was felt to be a 'redo' button by four respondents and a rotation control by three more, although the direction confused them. Icon 10 was related with confidence to a media 'play' control by 10 respondents and YouTube® was cited correctly by three, showing the influence of popular social media software.

Icon 11 was identified as a 'pause' control by nine respondents and YouTube® was correctly cited in one case. Icon 12 caused confusion, with respondents seeing it as like icons as diverse as a Wii® controller and an icon from photograph viewing 'apps'. Icon 13 was related to 'map apps' by two respondents and as a 'zoom' control by two more. Icon 14 is the opposite of Icon 8, and five respondents gave it the opposite meaning. Icon 15 (the opposite of Icon 9) was perceived correctly by one respondent as a rotation symbol. Icon 16 was understood as a common 'information' icon by eight respondents, but three of them were confused about whether it was general or specific information.

Icon 17 was viewed with confidence as an 'email' button by thirteen respondents, although no-one deduced its secondary meaning. Icon 18 also caused confusion, some respondents seeing it as a 'no entry' sign. Icon 19 was also seen to resemble icons used in several different common applications. Only two respondents saw it correctly as a navigation control who felt that it was like a control from Google Street View® or an X-box® icon. Icon 20 was recognized as a map pointer as used in Google Maps ® by ten respondents with confidence, but most failed to identify it as a 'jump' device. Icon 21 created the most confusion, as five respondents perceived similarities to other interfaces (e.g. Google Street View®) and one felt that it was like the 'collapse' control on a pull-down menu.

Question 3. *'Does grouping icons in tool bars make their meaning clearer?*

Seven participants (33.3% of the sample) felt that grouping the icons into tool bars had *not* made their meaning clearer. A typical response was, 'The tool bars do not make any difference…you look (locate) and use the *icon* you require, not the tool bar'. For those respondents who did perceive a positive difference, four significant themes emerge from the analysis:

- **Position** - where a toolbar is placed on the screen (e.g. right or left) suggests navigation in either direction, while the center or top of the screen suggests a more general use;
- **Difference** - icons need to be clearly distinguishable from other icons in the same set;
- **Proximity** - the closeness of a toolbar to an item on the screen (e.g. a painting or a doorway) suggests the meaning of the icon and its intended purpose;
- **Consistency** - the icons in a tool bar should perform functions regularly so that they are learned and understood more easily (e.g. the main tool bar is used more frequently and consistently than the left or right tool bars);
- **Association** - links between groups of icons in a tool bar suggest their use and meaning, which may be transferred over from other applications using similar icons.

The Thematic Analysis shows that an understanding of the meaning of the icons in a tool bar relates strongly to their position in relation to other items on the screen. For the Main Tool Bar ten out of 15 respondents felt correctly that its position (at the top center of every screen) suggested that the icons had a general purpose. For the Left Tool Bar seven out of 12 respondents and the also seven out of 15 respondents for the Right Tool Bar were correct in feeling that the position of the respective tool bars indicated that they operated on the object being viewed (e.g. zoom in or out, rotate left or right, etc.). The other themes were less strongly indicated but with the Right Tool Bar, association (e.g. with other icons in the tool bar) indicated a specific application in five out of 14 responses. This tool bar contained both navigation and information icons and some respondents could not distinguish between them. This appears to reinforce research [40], which suggests that icons need to maintain 'difference'. An icon needs to be clearly distinguished from other icons in the same tool bar and be close semantically to its own function while maintaining as great a semantic distance as possible from the other icons.

For the Main Tool Bar 'consistency' and 'association' were both cited as indications of meaning in two out of 15 responses, whereas 'proximity' and 'association' (e.g. with other icons in the tool bar) were cited in only two out of 12 responses in relation to the Left Tool Bar and in four out of 14 responses to the Right Tool Bar. One respondent offered the comment, 'I think it's a good idea to make the right-hand tool bar look like the left one (only [including] navigation icons) and move the other icons (information ones) in the right tool bar to the top [Main Tool Bar]'. An example of 'association' occurs in the response of one user, who associated Icon 16 'Information' with Icon 17 'Send email', which appear together on the right tool bar, assuming the 'i' symbol led to an address book while the 'envelope' referred to sending the email. Another suggestion was to remove Icon 6 and Icon 7 that vary the image size from the Right Tool Bar and place them on the painting itself, like Icon 18 that closes a window.

12. Discussion of findings from Test One and Test Two

From the results of this study it is possible to make certain observations. Icons that resemble their intended function more closely (i.e. have a close sematic distance) tended to have a higher IRR score both 'out of context' and 'in context'. It can be concluded that this is because less prior learning or familiarity is needed for users to understand their meaning. As computer icons are not 'standardized' as are warning signs through the ISO [26-27] icon designers' adaptation of the same or similar icons for different purposes can create misinterpretation.

Theory suggests that when planning an interface, icon designers have a conceptual model of the way in which the icons will be used [41] based on their training and experience. The users of the interface, on the other hand, will have a mental model of the icons' meaning based on their knowledge, cultural back ground and familiarity [41]. The importance of matching these models is demonstrated by the confusion caused in the tests by 'familiar' icons whose functions differed from users' expectations. The IRT 'out of context' (Test One) showed that 33.3% of the icons were

clearly identifiable to users (see Table 9). Icon 4 was confused with a 'gear cog' for adjusting system settings. Icons 9 and 15 were too similar to icons used differently in other applications. Icon 21 had been encountered with touch screens for 'swiping', but not for navigating between displays, having an adverse effect on usability through a "*lack of conformity with user expectations*" [42].

13. Conclusions

The original pilot study suggested that there was a problem with icon recognition, even to expert, qualified computer users [1]. This prompted that further research needed to be done into the phenomenon. The research for this paper is therefore an extended study based on an expanded sample of 21 computer users with different levels of competence, ages, educational attainments and spheres of employment. It is felt that this sample represents typical users of a virtual art gallery.

The extended research project set out to evaluate a set of randomly-chosen icons that carry out the action, information and navigation functions in a virtual gallery interface. A combination of quantitative and qualitative techniques was employed to add depth to the data analysis, while avoiding the use of complex statistics. The study is therefore based on an established method of Icon Recognition Testing (IRT) examining 21 icons from a 'real world' virtual gallery. However, this study is original, as it combines tests in and out of context and draws a comparison between them. An additional innovation is the combination of qualitative textual and thematic analysis (Sections 10 and 11) to establish reasons for the users' interpretation of the icons' meaning. This adds considerably to the contribution of the research. The findings are useful to interface designers and academics alike, by offering advice and by prompting further research. Conclusions for virtual interface design are drawn from the results of the research under the following headings:

13.1. Familiarity helps – people get to know icons

The IT industry is a long way from adopting standards equivalent to those for warning and traffic signs. However, the study shows that familiarity can aid users in recognizing their meaning in different contexts. The consistent use of familiar icons for their expected function is therefore important. When an interface is used regularly (e.g. a word processing package) users gain familiarity with the icons' function, even by its position, without having to decode its meaning. However, individuals tend to visit a virtual gallery relatively infrequently and are less likely to gain familiarity with the icons. The low IRR scores of icons that are 'custom made' for the 'Artweb.com' interface (e.g. Icons 4, 9 and 15) would appear to support this conclusion.

13.2. Abstraction is useful – but should be controlled

This paper begins by discussing concreteness and abstraction in icon design. The research shows that augmenting an icon with text (e.g. Icon 1) assisted the users in understanding its meaning. Therefore, adding more visual detail to the icons (i.e. making them more concrete) may reduce ambiguity. However, it may initially take longer for users to process mentally [16] and could interfere with their enjoyment and detract from the virtual experience. The balance between abstraction and concreteness should be an important consideration for interface designers.

13.3. Icons should be 'audited' regularly

The extended study prompts the recommendation that icon recognition testing should be carried out regularly as a part of an interface design 'audit' to ensure that the icons are continuing to fulfil their intended purpose. The study suggests that after such an audit, icons classed as 'identifiable' should be maintained in their present form. It is further suggested that icons classed as 'mediocre' could be modified economically to be more effective by making them more concrete or sufficiently different from icons used for other functions. However, icons classed as 'vague' should be redesigned completely or replaced, taking into account the recommendations offered in Sections 13.1. 13.2 and 13.3. It is suggested that some icons in this category may be replaced by familiar icons from other software packages that have passed ISO benchmark tests (e.g. MS Word®), subject to legal approval.

13.4. Interface designers need to understand user profiles

The results demonstrate that when designing icons for a virtual interface (in this case a virtual art gallery) it is important that the designer's conceptual model closely matches the users' mental model. Norman [41] explains that the interface designer does not communicate directly with the user, but through the 'system image', which is developed from the designers' own conceptual views and understanding of the nature and purpose of the interface. The users subsequently form a mental model based on their own understanding and interpretation of the system image, influenced by their beliefs, experience and prior knowledge (i.e. their user profile). A match between the conceptual model, the system image and the user profile should result in an enhanced user experience, so virtual interface designers should capture user profiles to adapt the interface to the user's requirements.

14. Limitations of the research

Importantly, this study has its limitations. The IRT focussed on evaluating icons with different functions taken from the same interface (i.e. Artweb.com). The study by Ferreira *et al.* [20] compared icons from different interfaces with the same function. In both Ferreira's and this research, the tests were limited to identifying the icons' meaning using paper-based tests. A more sophisticated and comprehensive icon recognition test could be done with technology that would record more information about the users' intuitive responses and 'thinking time' (e.g. interactive MS PowerPoint with key logging). The tests could be extended so that the participants could compare different virtual interfaces.

15. Suggestions for future research

Many of the virtual gallery interfaces identified in the secondary research currently offer a 'one size fits all' approach to icon design. It is suggested that more needs to be known about the potential for user profiling in virtual interface design, perhaps using an ontology engineering approach. Methods of capturing this profile need to be non-invasive if the user experience is not to be compromised. There is the potential for exploring methods such gamification as a way of capturing users' profiles and preferences. A variety of frameworks exist to enable designers to do this [43].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

We thank the participants who evaluated the icons and provided detailed interpretations. We also acknowledge the expertise of the designers of the German virtual art gallery website for providing a rich and useful source of enjoyment to art lovers and material for researchers into human-computer interaction.

References

[1] D.E. Ashe, W.A. Eardley and B.D. Fletcher, "e-Tourism and Culture through Virtual Art Galleries: a Pilot Study of the Usability of an Interface" in 4th International Conference on Information Management (ICIM), Oxford, U.K. 2018. http://dx.doi.org/10.1109/INFOMAN.2018.8392834.

[2] Smithsonian Natural History Museum virtual tour. 2018. http://naturalhistory.si.edu/panoramas/.

[3] Louvre Online Tours. 2018. http://www.louvre.fr/en/visites-en-ligne.

[4] Virtual Tour of the Oxford University Museum of Natural History. 2018. http://www.chem.ox.ac.uk/oxfordtour/universitymuseum.

[5] Virtualfreesites. Virtual Tours of Museums: Exhibits and Points of Special Interest: Virtual Tours. 2018. htpp://www.virtualfreesites.com/museums.html.

[6] W.A. Eardley, D.E. Ashe and B.D. Fletcher, "An Ontology Engineering Approach to User Profiling for Virtual Tours of Museums and Galleries" Int. J Knowl. Eng., **2**(2), 85-91, 2016. http://dx.doi.org/10.18178/ijke.2016.2.2.058.

[7] D.J. Mayhew, Principles and Guidelines in Software User Interface Design. New Jersey: Prentice Hall, Inc. 1992

[8] P.B. Anderson, A semiotic approach to programming. In: The computer as medium. Cambridge, England: Cambridge University Press, 16-67. 1994

[9] J. Ferreira, P. Barr & J. Noble, "The semiotics of user interface redesign" in Proceedings of the 6th Australian User Interface Conference (AUIC '05), **40** (4) Sidney, Australia, 47-53. 2005.

[10] R. Yan, "Icon Design Study in Computer Interface." Pro. Eng, **15**, 3134-3138, 2011, http://dx.doi.org/10.1016/j.proeng.2011.08.588.

[11] H.I. Cheng and P.E. Patterson, "Iconic hyperlinks on e-commerce websites" J. Appl. Ergon, **38** (1), 65-69, 2007, https://doi.org/10.1016/j.apergo.2006.01.007.

[12] B. Merdenyan, O. Kocyigit, R. Bidar, O. Cikrikcili and Y.B. Salman, "Icon and User Interface Design for Mobile Banking Applications" in Proceedings of the 4th International Conference on Advances in Computing and Information Technology (ACITY '14), **4**(2), 55-59, 2014, https://dx.doi.org/10.3850/978-981-07-8859-9_18.

[13] G. Bhutar, R. Poovaiah, D. Katre and S. Karmarkar, "Semiotic Analysis combined with Usability and Ergonomic Testing for Evaluation of Icons in Medical User Interface" in Proceedings of the 3rd International Conference on Human Computer Interaction (IndiaHCI '11), Bangalore, India, 57-67. 2011. http://dl.acm.org/citation.cfm?id=2407804.

[14] P. Barr, J. Noble, and R. Biddle. "Icons R Icons" in Proceedings of the Fourth Australasian User Interface Conference on User Interfaces, 18 (AUIC '03), Darlinghurst, Australia: Australian Computer Society, Inc., 25-32. 2003, https://dl.acm.org/citation.cfm?id=820093

[15] C. Gatsou, A. Politis and D. Zevgolis. The Importance of Mobile Interface Icons on User Interaction. [Online] Int. J. Comput. Sci. Appl., **9** (3), 92-107. 2012. http://www.tmrfindia.org/ijcsa/v9i37.pdf.

[16] L. Nadin. Interface design and evaluation - Semiotic implications. In: H.R. Hartson and D. Hix (Eds) Advances in human-computer interaction, **2**. Norwood, NJ, USA: Ablex Publishing Corp.1988.

[17] R. Scalisi. "A semiotic communication model for interface design" in Proceedings of the 1st International Conference on Computational Semiotics for Games and New Media (COSIGN2001). Netherlands, Amsterdam. 10-12 Sept. 2001. http://www.cosignconference.org/conference/2001/

[18] S. Isherwood. "Graphics and Semantics: The Relationship between What Is Seen and What Is Meant in Icon Design" in Proceedings of the 8th International Conference on Engineering Psychology and Cognitive Ergonomics, Held as part of HCI International, San Diego, California, USA. July 2009. https://doi.org/10.1007/978-3-642-02728-4_21.

[19] R. Arnheim, Visual thinking. London: Faber. 1969.

[20] J. Ferreira, J., Noble and R. Biddle, "A case for iconic icons" in Proceedings of the 7th Australasian User Interface Conference (AUIC '06), Hobart, Tasmania, Australia: Australian Computer Society, Inc., **50**, 97-100. Jan. 2006. https://dl.acm.org/citation.cfm?id=1151771

[21] Y.B. Salman, H-I. Cheng and P.E Patterson. "Icon and user interface design for emergency medical information systems: A case study" Int. J. Med. Inf., **81**(1). 29-35, 2012. http://doi.org/10.1016/j.ijmedinf.2011.08.005.

[22] S. Schröder and M. Ziefe, "Making a completely icon-based menu in mobile devices to become true: a user-centred design approach for its development" in Proceedings of the 10th International Conference on human computer interaction with mobile devices and services, Amsterdam, The Netherlands: ACM, 137–146, Sep. 2008. http://doi.org/10.1145/1409240.1409256.

[23] J. Nielsen and D. Sano, "SunWeb: User interface Design for Sun Microsystem's Internal Web". Comp. Nets. ISDN Sys, **28**(1), 179-188., 1995. https://doi.org/10.1016/0169-7552(95)00109-7.

[24] J.J. Foster. Graphical Symbol: Test methods for judged comprehensibility and for comprehension. ISO Bulletin, 11-13, December 2001. http://hablamosjuntos.org/signage/PDF/graphicsymbols0112.pdf. 2001.

[25] S. Ghayas, S. Sulaiman, M. Khan and J. Jaafar. The effects of icon characteristics on users' perception. In Advances in Visual Informatics. Switzerland. Springer International Publishing, 652-663. 2013.

[26] ISO. ISO 9186:2011. Graphical symbols - Test methods for judged comprehensibility and for comprehension. International Organization for Standardization. Geneva, Switzerland, 2001.

[27] ISO. ISO 3864:1984. Safety colours and safety signs. International Organization for Standardization. Geneva, Switzerland, 1984

[28] W.C. Howell and A. H. Fuchs, "Population stereotypy in code design", J. Org. Behav. Hum. Perf., **3**(3), 310–339, 1968, https://doi.org/10.1016/0030-5073(68)90012-3.

[29] E. Kaasinen, J. Luoma, M. Penttinen, T. Petäkoski-Hult and R. Södergård. Key Usability and Ethical Issues in the NAVI programme (KEN). 2001. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.4971&rep=rep1&type=pdf.

[30] B. Martin and F. Ringham, Key terms in semiotics. New York: Continuum International Publishing Group Ltd.,2005.

[31] P.R. Desai, P.N., Desai., K.D. Ajmera and K. Mehta, K., "A review paper on oculus rift - A virtual reality headset". Int. J. of Eng. Trend. & Tech. (IJETT), **13** (4), 175-179, 2014. https://doi.org/10.14445/22315381/IJETT-V13P237.

[32] D. Biella, D., Sacher, B. Weyers, W. Luther and N. Baloian, "Metaphorical Design of Virtual Museums and Laboratories: First Insights" in International Conference on Ubiquitous Computing and Ambient Intelligence, 427-438. Springer, Cham, Dec. 2015.

[33] Statista.com. Museum attendance in England by age 2012-2017 UK survey. 2018. https://www.statista.com/statistics/418323/museum-galery-attendance-uk-england-by-age/.

[34] M.N. Islam and H. Bouwman, "Towards user-intuitive web interface design and evaluation: a semiotic framework" Int. J. Hum-Comput. St, **86**, 121-137, 2016, https://doi.org/10.1016/j.ijhcs.2015.10.003.

[35] L. Zakowska, "The effect of environmental and design parameters on subjective road safety - a case study in Poland" Safety Sci, **19** (2-3). 227-234, 1995. https://doi.org/10.1016/0925-7535(94)00023-V.

[36] E. Duarte, F. Rebelo, J. Teles, and M.S. Wogalter, "Safety sign comprehension by students, adult workers and disabled persons with cerebral palsy" Safety Sci. **62**, 175-186, 2014, https://doi.org/10.1016/j.ssci.2013.08.007.

[37] S. Rosenbaum and J. Bugental, "Measuring the Success of Visual Communication in User Interfaces" Soc Tech Commun's Technical Communication, **45** (4). 517-528, Nov. 1998.

[38] ISO. ISO 3864:2016 International standard for safety colours and safety signs. International Organization for Standardization. Geneva, Switzerland, 2016.

[39] S. McDougall and M. Curry, "More than just a picture: Icon interpretation in context" in Proceedings of 1st International Workshop on Coping with Complexity. University of Bath, 73. September. 2004. http://www.academia.edu/download/3251711/Past_Designs_As_a_Way_of_Coping_With_Complexity.pdf#page=73

[40] J.M. Silvennoinen, T. Kujala and J.P. Jokinen, "Semantic distance as a critical factor in icon design for in-car infotainment systems". Appl. Ergon., **65**, 369-381, 2017. https://doi.org/10.1016/j.apergo.2017.07.014.

[41] D.A. Norman "Some observations on mental models" in D. Gentner and A. Stevens (Eds.) Mental Models. Hillsdale, N.J., Erlbaum Associates. 1983.

[42] D. Oswald, "Dynamic sense-making in use processes of digital products". in 5th International Congress of the International Association of Societies of Design Research (ASDR), Tokyo. Aug. 2013. http://www.david-oswald.de/downloads/IASDR2013_oswald-interface-semiotics.pdf.

APPENDIX A

Effect of context on users' icon recognition rate and overall average IRR%

User	*Average IRR Score per User*										
	Gender	Age range	Job	Education	Skill Level	Out of context		In context		Change to IRR + or –	Overall Average
						Score/ Max. score	Percentage	Score/ Max. score	Percentage		
1	Male	26 - 33	Student	Bachelors	Adv.	20/42	47.6%	32/42	76.2%	+ 28.6%	**61.9%**
2	Male	26 - 33	Student	Bachelors	Adv.	28/42	66.7%	34/42	81.0%	+ 14.3%	**73.9%**
3	Male	42 - 49	Employed	Doctors	Adv.	25/42	59.5%	29/42	69.0%	+ 9.5%	**64.3%**
4	Female	60 - 69	Retired	College	Basic	15/42	35.7%	21/42	50.0%	+ 14.3%	**42.9%**
5	Female	26 - 33	Employed	Bachelors	Inter.	27/42	64.3%	27/42	64.3%	**0.0%**	**64.3%**
6	Male	42 - 49	Employed	College	Inter.	19/42	45.2%	25/42	59.5%	+ 14.3%	**52.4%**
7	Male	18 - 25	Student	College	Inter.	23/42	54.8%	32/42	76.2%	+ 21.4%	**65.5%**
8	Female	18 - 25	Employed	College	Basic	16/42	38.1%	26/42	61.9%	+ 23.8%	**50.0%**
9	Female	18 - 25	Employed	College	Basic	18/42	42.9%	23/42	54.8%	+ 11.9%	**48.9%**
10	Female	18 - 25	Student	Masters	Adv.	16/42	38.1%	25/42	59.5%	+ 21.4%	**48.8%**
11	Male	26 - 33	Employed	Masters	Adv.	27/42	64.3%	36/42	85.7%	+ 21.4%	**75.0%**
12	Female	18 - 25	Employed	School	Inter.	17/42	40.5%	31/42	73.8%	+ 33.3%	**57.2%**
13	Male	34 - 41	Home maker	College	Basic	27/42	64.3%	29/42	69.0%	+ 4.7%	**66.7%**
14	Female	50 - 59	Employed	Bachelors	Adv.	28/42	66.7%	30/42	71.4%	+ 4.7%	**69.1%**
15	Female	50 - 59	Employed	Bachelors	Adv.	12/42	28.6%	19/42	45.2%	+ 16.6%	**36.9%**
16	Female	34 - 41	Employed	Bachelors	Adv.	15/42	35.7%	25/42	59.5%	+ 23.8%	**47.6%**
17	Female	18 - 25	Employed	College	Basic	16/42	38.1%	31/42	73.8%	+ 35.7%	**56.0%**
18	Female	60 - 69	Employed	College	Basic	13/42	31.0%	19/42	45.2%	+ 14.2%	**38.1%**
19	Male	26 - 33	Student	Masters	Inter.	14/42	33.3%	16/42	38.1%	+ 4.8%	**35.7%**
20	Male	34 - 41	Student	Masters	Adv.	28/42	66.7%	29/42	69.0%	+ 2.3%	**67.9%**
21	Male	26 - 33	Employed	Masters	Inter.	29/42	69.0%	35/42	83.3%	+ 14.3%	**76.2%**
										Average of overall averages	**57.1%**

User icon recognition rate % = (Score / Total possible score) * 100.
Difference in IRR % = In context IRR % – out of context IRR%
* Change in IRR% **in red** = negative value and **in green** = positive value.
Average of overall averages (Column 7) = sum of all user overall averages / number of users
Overall averages **above** the average of overall averages are underlined

APPENDIX B

Icon recognition rate for each icon 'out of context' and 'in context' showing difference + or -

Icon			Out of context Score/Max score	Out of context IRR %	In context Score/Max. score	In context IRR %	Difference - out of context and in context
No.	**Image**	**Purpose**					
1		Action	33/42	**78.6%**	35/42	**83.3%**	+ 4.7%
2		Action	28/42	**66.7%**	33/42	**78.6%**	+ 11.9%
3		Information	22/42	**52.4%**	30/42	**71.4%**	+ 19.0%
4		Navigation	0/42	**0%**	5/42	**11.9%**	+ 11.9%
5		Information	17/42	**40.5%**	21/42	**50.0%**	+ 9.5%
6		Action	23/42	**54.8%**	32/42	**76.2%**	+ 21.4%
7		Action	21/42	**50.0%**	28/42	**66.7%**	+ 16.7%
8		Navigation	25/42	**60.0%**	35/42	**83.3%**	+ 23.3%
9		Navigation	7/42	**16.7%**	4/42	**9.5%**	- 7.2%
10		Action	28/42	**66.7%**	31/42	**73.8%**	+ 7.1%
11		Action	30/42	**71.4%**	38/42	**90.5%**	+ 19.1%
12		Action	17/42	**40.5%**	28/42	**66.7%**	+ 26.2%
13		Action	25/42	**60.0%**	30/42	**71.4%**	+ 11.4%
14		Navigation	25/42	**60.0%**	36/42	**85.7%**	+ 25.7%
15		Navigation	9/42	**21.4%**	3/42	**7.1%**	- 14.3%
16		Information	24/42	**57.1%**	33/42	**78.6%**	+ 21.5%
17		Information	23/42	**54.8%**	29/42	**69.0%**	+ 14.2%
18		Action	18/42	**42.9%**	34/42	**81.0%**	+ 38.1%
19		Navigation	33/42	**78.6%**	38/42	**90.5%**	+ 11.9%
20		Navigation	16/42	**38.1%**	25/42	**60.0%**	+ 21.9 %
21		Navigation	9/42	**21.4%**	24/42	**57.1%**	+ 19.0%

Icon recognition rate % = (Score / Total possible score) x 100.
Difference in IRR % = In context IRR % – out of context IRR%
* Icon type **in bold** = dominant type of icon in sign
IRR score underlined would pass ISO 3864-2:2016 level of 66.7%

Identifiable (60% - 100%)	**Mediocre** (59% - 30%)	**Vague** (29 - 0%)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 298-306 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

The Model Development of an Effective Triggering System of Production Kanban Size towards Just-In-Time (JIT) Production

Mohd Norzaimi Che Ani[*,1], Shahrul Kamaruddin[2], Ishak Abdul Azid[3]

[1]*Manufacturing Section, Universiti Kuala Lumpur – Malaysian Spanish Institute, 09000, Kedah, Malaysia*

[2]*Mechanical Department, Universiti Teknologi Petronas, 32610, Perak, Malaysia*

[3]*Mechanical Section, Universiti Kuala Lumpur–Malaysian Spanish Institute, 09000, Kedah, Malaysia*

ARTICLE INFO

Article history:
Received: 30 July, 2018
Accepted: 02 October, 2018
Online: 12 October, 2018

Keywords:
Effective Kanban size
Triggering System
Dual feeders
Just-in-Time (JIT)
Waiting time

ABSTRACT

The contents of this article consist of an extension original work presented at the 4th International Conference on Control, Automation and Robotics (ICCAR 2018) and aims to develop the systematic model of an effective triggering system of production Kanban size towards Just-In-Time (JIT) production system. The developed model was introduced based on the philosophy of continuous improvement program by selection of case study industry in automotive manufacturing floor and it has been carried out through the development of an effective material Kanban size model using triggering system towards achieving a JIT production system. This developed model was analysed and verified by implementing through three phases which are the investigation of the Kanban size, analysis the production idling time and analysis of the time management for material replenishment process. The implementation was carried out to ensure the main objectives of the developed model in minimizing the waste of waiting through effective Kanban size for production system will be achieved by focusing on the communication between material incoming warehouse and production floor. The significant results were obtained during the verification process where the dual material feeders were identified for every process in the production floor to ensure the effectiveness of the production system. Throughout the implementation of the developed model, the production losses time caused by miscommunication between the production floor and warehouse had been identified and it was successfully minimized and the waiting time was reduced by 36.64%.

1. Introduction

In the current revolution of the production system, the crucial issues of efficiency and effectiveness of manufacturing flow are the main challenges in meeting the demands as expected by the customer. Jasti and Kodali [1] have stated in their study, in encountering this challenges, the main requirement in fulfilling customer expectations is industrial practitioners required to accelerate their corrective action by performing continuous improvement program to survive and increase the competitiveness. One of the effective approaches practiced by automotive industry is Lean Manufacturing (LM) concept which is pioneered by Toyota Production System (TPS). The ultimate goals of LM concept are eliminated the wastes of operation in between process to process in manufacturing plant and ensure the smoothness of the production process to ensure the process flow are streamlined and efficient [2]. The LM concept was industrialized through the techniques of process improvement program by introduced a variety number of tools and proper methods to ensure the continuous improvements culture is adopted among the employees. TPS is targeted to eliminating or minimizing all type of identified wastes and smoothing the production flow by utilizing the LM philosophy known as Just-in-Time (JIT) and Jidoka or also known as production leveling [3]. JIT production system developed by TPS is an efficient concept with the main aim to minimize the production lead times within operation of production floor as well as quick response times in solving any arising issues between process to process, or communication between material supply chain system [4]. The main purpose of JIT system is to increase the smoothness of production flow and minimizing the overall

* Mohd Norzaimi Che Ani & Email: mnorzaimi@unikl.edu.my

costs of manufacturing operation covering minimization of the reject rate between process to process, optimizing the resources of production floor including space allocation for required processes, movement of the direct and indirect workers, standardization of the raw material inventory and synchronization of the internal and external transportation [5].

As introduced by TPS, LM concept focusing on the seven types of production wastes known as transportation, inventory, motion, waiting, over-production, over-processing, and defects. In the group of seven types of wastes, the most critical waste in influencing the efficiency and effectiveness of production flow is managing the inventory. Inventory means the management of raw materials from the warehouse into production floor or semi-finished products between process to process in production floor chain system, normally called as Work-in-Progress (WIP). Ineffective coordination and controlling of the WIP in the production floor causing the number of inventory will keep increasing and queues of the raw material prior transfer into production floor causing the waste of waiting in between process-to-process. As introduced by TPS in solving this issue, the LM concept has been introduced through implementation of the Kanban system. Kanban is a Japanese word was introduced in the scheduling system of raw material management for LM concept and JIT production system and it works as a management and controlling the inventory system either between production floor and raw material warehouse or management the internal supply chain which is controlling the WIP between process to process in production floor [6, 7]. As current practice, most of the automotive industries was implementing the Kanban system and the main benefits was obtained is ability to standardize an optimum limit or quantity of the inventory or WIP and preventing tasks of overloading during performing production processes, or also known as minimizing the 3M (Muda, Mura, Muri) [8]. LM concept well accepted and recognized by most of the industrial practitioners in the worldwide based on fruitful accomplishment in TPS especially controlling the inventory through one piece flow activity especially in automotive industries. However, the implementation of the one piece flow activity in di different types of industries is impossible as implemented by TPS because of several factors such as distance and location between suppliers and customers, facilities of the transportation system, a designed system of production flow and the availability of the facilities inside production floor. In tackling this issues, an effective WIP system through standardization of Kanban size was introduced with the main purpose to ensure effective management of holding inventory and minimize the waiting time in the production floor [9, 10].

In the current scenario, as practiced by the majority of the industrial practitioners in improving the efficiency and effectiveness of production system through implementation the Kanban system, but unstandardized numbers of inventory due to multiple products in production floor is required and it was caused miscommunication between production floor and incoming material warehouse due to inefficient inventory management system. Even though Kanban system widely applied in automotive industries but the main issue facing by automotive industry is to every single unit of finished-good vehicle needs thousands number of raw materials or semi-finished product and dealing with the various level of suppliers [11]. Hence, a direction of this research paper, the implementation of the Kanban system using an effective number of feeders by standardizing Kanban size for in-house production floor has been developed and introduced. In ensuring the effectiveness of the developed Kanban system, the "triggering system" of management WIP was initiated to ensure production floor meeting the principles of JIT production system.

The focus of this paper is directed towards various literature that discussed the implementation of Kanban system in manufacturing system especially the inventory interaction between the incoming material warehouse and production floor. The first objective is to identify the critical success factor of achieving the optimum efficiency and effectiveness of production system by exploring the mechanism of the production flow. For a second objective, even though the Kanban system has been successful in manufacturing industries and widely recognized, but as the issues in effective communication between an incoming material warehouse and production floor are usually overlooked, the effective triggering system production Kanban size is introduced. In this paper, Section 2 describes the concept of production system to identify the critical area of influencing production efficiency and effectiveness, Section 3 presents on the development of the Effective Triggering System Production Kanban Size model, Section 4 presents the results and discussion of model verification in selected case study, and Section 5 concludes the overall results of the developed model.

2. The Concept of Production System

In this section, the current concept of the production system in various manufacturing industries of the production system will be discussed and explored. Since the production system is the main subject in manufacturing industries, the production efficiency and effectiveness must be closely monitored and performed to ensure the organizations meet the customer expectation and become competitive with other competitors. As indicated in the era of global competitiveness, the manufacturing organizations are enduring enormous pressure from their customers and competitors [12]. All these factors have given means to integrate the effective system of production flow with the complete production process (starting from the suppliers to the delivery to the customer) [13].

Expanding the inductive research approach, the attention has been focused on the flow of the production system according to the general factory set-up from various industrial sectors. That means, practically, the production system can be segregated into three sections; incoming raw materials, production processes and outgoing finished good product. In general interpretation, the production system can be summarized as input, process, and output. The input consists of incoming raw material, known as downstream area, which requires multiple numbers of tasks to ensure supplying the right parts at the right time into the process. The process in the production floor involves multiple numbers of sub-processes of semi-finished product and the process is continued until product completion. The output consists of inspection, packaging and shipping the good product to the customer based on the order in outgoing section.

2.1. Key Elements of Production System

As the production system can be generally divided into three sections, the key elements of the production system involving all

these sections. For input, the management of the raw material begins with the preparation of the raw materials either semi-finished product from sub-contractor or the child part from suppliers. A verification process of the raw materials by OEM normally goes through acceptance sampling analysis to ensure that raw materials meet the specification [14]. Acceptance sampling involves the process of inspection and testing of the product. A sample of goods is taken randomly from the lot or batch and a decision is made whether to accept the raw materials from suppliers based on the results of acceptance sampling and it will determine whether a batch of the raw material of goods should be accepted or rejected [15]. Once the quality of raw material is confirmed, it will be stored in warehouse storage area for the inventory purpose. The raw material will then be retrieved from the storage area, and the assigned worker will perform the preparation activity to ensure the specification quality and quantity according to the order from production, and this whole process is known as an upstream process. The raw materials will be temporarily parked at designed area, normally called kitting area or Work-in-Progress (WIP) area, with the purpose of the replenishment process. At this stage, the raw materials are now ready for use in production processes.

The next element is process, also known as midstream process, which means the raw materials will be processed from raw materials or semi-finished products. The production process is concerned with transforming a range of inputs into outputs as required according to the customer [16]. In the production processes, normally the Statistical Process Control (SPC) will be applied to ensure the finished product meets the specification. SPC is a statistical technique used for measuring, analyzing and controlling process variation in order to prevent defects or non-conformance in products. SPC involves inspecting a random sample of the output from a process and it will decide whether the process is producing products with characteristics that fall within a predetermined range [17]. Abnormality results from the SPC process will lead to the process improvement activity in order to solve the issue. After production processes are completed, and then the finished product will be transferred to the next section which is the output.

In the output stage, also known as a downstream process, the product will involve with the specific package which are packaging and shipping [18]. The finished product will be packed according to the customer expectation and at the same time, the assigned worker will perform the visual inspection to ensure that the finished product is free from defects. Some of the issues might arise at this stage such as cosmetic issue due to handling process and lead time. After completing the packaging process, then the finished product will be transferred into the transportation system for the shipping process. The production system is considered completed once it reaches the stage of shipping.

2.2. Issues pertaining to the elements in Production System

In the current global scenario, the competitiveness in the manufacturing industries keeps increasing as the customers ask for the minimization of the operation costs and increased quality of the product at the same time [19]. In order to ensure meeting this situation, the efficiency and effectiveness of production system play main roles in improving the production system and at the same time minimizing the cost of operation. To achieve these objectives, the whole chain of the production system, upstream, midstream and downstream, should the improved. This is also highlighted by Lean Manufacturing (LM) approach, where the process improvement should be studied from suppliers to customers [20]. Since the upstream section is the predecessor process for the production process, it should be prioritized to achieve optimum efficiency of the production system. Even though the upstream process is typically seen to play a minor role in the production system, it affects the lead time, part shortage, part mixing or product failure, which leads to inefficient production. Thus, the upstream process must achieve high efficiency, reliability and traceability to ensure the parts transferred into the production process meet the expectation.

In achieving high efficiency of communication for the upstream process, the synchronization interaction between warehouse and production must be effective in term of planning daily production required, quality requirement and time management of raw material replenishment process. Usually in achieving effective synchronize communication, the process of internal supply chain between warehouse management and production floor will be connected using inventory control system [21]. The chain of the production process was started with warehousing because it is the first section in the manufacturing process which is responsible controls physical of the receiving material and it performs the material verification tasks during the process of receiving from suppliers or vendors, incoming material sampling analysis and segregating the raw material before transferring into the production floor. The management of the warehouse must know the physical shape of the raw material and also know all information of the time and place of supplying raw material into production floor. The process of monitoring and verification of the raw material inventory is about the measurement of the critical parameters and quality inspections for achieving an effective management of the inventory [22]. Inefficient inventory management in the upstream process will be caused interruption of the whole production flow and it can lead to high stoppage time, increasing the production lead times, congested queues of transferring raw materials, smoothness of product flow and increasing costly operation setups [23].

3. The Development of Effective Triggering System of Production Kanban Size Model

In order to ensure achieving the optimum efficiency and effectiveness of production system, an effective triggering system for Kanban management system was developed and introduced in this research-based philosophy of continuous improvement program through three phases. The first phase required analysis of Kanban size then follows by the second phase consists of the investigation of the production idling time using queuing analysis and third phase the material preparation ‘triggering system’ was introduced. The framework development of an effective triggering system for Kanban management system is illustrated in Figure 1.

Based on the framework illustrated in Figure 1, in the first phase, the analysis of the production Kanban size was properly carried out in ensuring achieve the optimization of the material replenishment cycle through minimizing the production losses time during raw activity of raw material replenishment. During

executing this phase, the actual size of production lot size is required to be identified to ensure determination of the suitable number of production lot size achieving the optimum solution. The number of the production lot size considered same from day-to-day production operation based on long-term production planning but the quantity of the raw material or semi-finished products from suppliers would be different due to the production requirement as daily planned by production planner. In the determination of an effective production Kanban quantity, the reliable information on the individual process cycle time from production floor and raw material preparation per unit of the finished-good product were needed to ensure the accuracy of calculating the effective Kanban size as formulated in Equation (1).

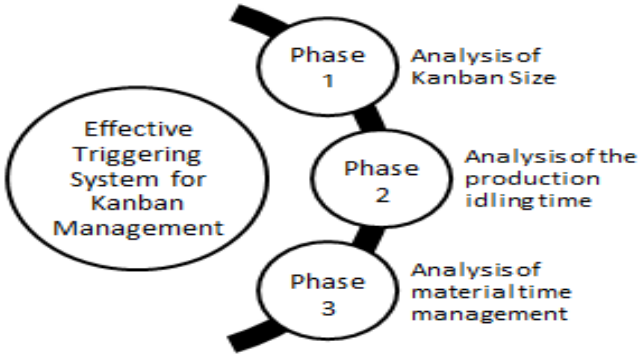


Figure 1. The framework development of effective triggering system for Kanban management system

$$\text{Effective Kanban size} = \frac{\text{Material Prepration Cycle Time}}{\text{Production Cycle Time}} \quad (1)$$

The Kanban size was identified to ensure planning of the upstream process in proper manner. Once the Kanban size was determined, the analysis of every motion performed by warehouse personnel for the upstream process will be analysed. The worker will be performed the assigned tasks to complete the material preparation process and time will be recorded. Since the warehouse activities and production floor normally set-up based on segregated building, the traveling time from the warehouse into the production floor must be considered in motion time study analysis of material preparation activities. This situation created the waste of transportation and the production process will be interrupted during the material replenishment process.

Once the Kanban size has been determined in Phase 1, then in the following phase which is Phase 2 the application of the queuing system theory will be used to analyse the production idling time for each process in production floor with the purpose to identify the wastes of production waiting time. The queuing theory model consisted of mathematical study based on the application of the probability concept to identify the production waiting time, or idling time-based on the queuing behavior of raw material management [24]. The application queuing theory model was considered in this research, because in the queuing analysis theory, the queuing behavior will be counted as raw material queues waiting for prior transfer into production floor is based on consideration of the production processing time and cycle of raw material replenishment from warehouse into production floor, and both of the parameters are then will be converted into hourly service rate, known as μ and hourly arrival rate, known as ʎ. The determination for both μ and ʎ were analysed based on Equations (2) and (3).

$$\text{Service rate, } \mu = \frac{1 \text{ hour}}{\text{mean service cycle time}} \quad (2)$$

$$\text{Arrival rate, } \lambda = \frac{1 \text{ hour}}{\text{mean arrial rate}} \quad (3)$$

In queuing system theory, a customize model of queuing behavior is constructed to determine the queue lengths (number of inventory) and the waiting time (production idling time) that can be predicted based on systematic analysis of queuing behavior either single queuing behavior model or multiple queuing behavior models. As practiced by most of the industries, normally production floor where the process-to-process in connected and receiving the raw material from the warehouse, the material preparation activity in warehouse considered as a supplier for production floor. Based on this condition the single queuing is more suitable for the further analysis. Single queueing theory model normally practices by using Kendall's notation in the form A/S/C where A describes the time arrivals to the queue, S the size of queues and C the number of servers or counters at the node [25]. Even though the single theory model required a single queues length and single server, but multiple situations of queuing behavior are possible to adopt. Based on available theories of algorithms for basic single theory model, several conditions have been identified such as basic single server, constant or undefined service time, finite queuing line and finite calling population. The selection of an application for each condition depending on the activities conducted in queuing behavior. Shortle *et al.* [26] were segregated the group of basic queuing theory as summarized in Table I.

Table 1: Type of queuing system

Theory of Algorithms	Queuing System	Queuing Behavior
Single server	Basic single server	Determined cycle time
Single server	Undefined or constants service time	Undefined or fixed cycle time
Single server	Finite queuing line	Finite input
Single server	Finite calling population	Finite server
Multiple server	Multiple servers waiting line	Multiple similar servers

Once a single model of the queuing system was determined, then detailed analysis using probability theory in queuing analysis system will be carried out. In queuing theory, generally five main parameters will be analysed which are the probability the server is idle (Po), the average of number of customer in the queuing system (L), the average number of customers in the waiting line (Lq), the average time a customer spends in the total queuing line (W) and the average time customer spends waiting in the queue to be served (Wq). As practiced in the production system, smoothness interaction between warehouse and production floor is required, the definition L and W was customized to suit with a production system. L was defined based on the number of production

processes waiting to be completed for material replenishment activity in production floor, while W was defined as an average time the production process waiting for completion of the material replenishment activity. Subsequently, the determination for both of the L and W will be based on the identified service rate, μ, and arrival rate, ʎ as mentioned earlier. Since this research focusing on the material replenishment process versus the production system, the most significant queuing behavior system was finite calling population as indicated in Table I. It was used in this research because the main constraints during material replenishment activity is the quantity of material feeder where is considered as terminal or server and the production processes considered as customer for the warehouse during replenishment activities. Furthermore, in ensuring the reliability of the determination of both L and W, the number of entire processes in production floor also considered. The determination for both L and W for finite calling population queuing system was formulated as Equation (4) until Equation (8).

$$Po = \frac{1}{[\sum_0^N \frac{N!}{(N-n)!}(\frac{\lambda}{\mu})^n]} \quad (4)$$

$$Lq = N - [\left(\frac{\lambda+\mu}{\lambda}\right)(1-Po)] \quad (5)$$

$$L = Lq + (1-Po) \quad (6)$$

$$Wq = \frac{Lq}{(N-L)\lambda} \quad (7)$$

$$W = Wq + \frac{1}{\mu} \quad (8)$$

Successful implementation and completion of Phase 2, the 'triggering system' for effective management of WIP will be developed in the third phase. In this phase, time management analysis of inventory planning was carried out in determination the cycle time of WIP 'triggering system' in ensuring sustainability of the developed production Kanban system towards achieving optimum efficiency and effectiveness of production system. Hence, in this phase, the raw material replenishment time from warehouse and cycle was determined based on the daily production scheduling and planned operation hours to ensure the effective preparation time of warehouse activity. The identification of the cycle for material replenishment activity will be determined based on planned working time per days such as a number of working shifts and the daily targeted output as per planned by production planning. Once the material replenishment cycle and time was identified, then the traffic management for movement of raw material activities in production floor could be identified to ensure smoothness of material replenishment process. This is important because the process of transferring the raw materials from the warehouse into production floor required high repetitive movements and traveling due segregated warehouse and production floor from different buildings. In this phase, the proper management of internal transportation system is required by implementing visual management system such as color coding and alarm system.

4. Results and Discussion of Model Development

The selected case study industry for this research is an automotive manufacturing industry which operating the process of assembling the car. The production system was designed using single production assembly layout with connected from process-to-process and the entire production floor and it was divided into five sections which are a trim section, chassis section, final assembly section, accessories assembly section, and final inspection section. The raw material warehouse was segregated from the different building which is next to the production floor building and the material preparation and acceptance sampling process were performed in the warehouse building and it will be transferred into the production floor based daily production planning. Each section in the production floor and warehouse management was managed by the different groups of management and it needs an efficient material preparation activity during establishment the smoothness of production system.

Since the processes of production floor in the selected case study was set-up based on five sections, one of the section will be selected as initial improvement area to be conducted to verify the developed model in this research. Even though five sections is required in assembling the car, but only three sections were considered and possible conducting the improvement activity of verification process, which are Trim section, Chassis section, and Final Assembly section. These three sections were identified as potential sections in influencing unsmooth production process due to interaction with the multiple types and quantity of raw materials from the warehouse. The other two sections which are Accessories Assembly and Final Inspection were not involved with replenishment of raw material because only conducting the process of checking and verification of finished-good product and it was considered as a stand-alone process. Based on the several criteria, Trim section was selected and in implementing the developed model as a case study for this research. The Trim section was selected according to two criteria which are this section considered as a predecessor for overall production chain system because it was set-up as first process and the failure or inefficient process of Trim section would be interrupted the entire production flow. Furthermore, which is the second criterion in the selection of Trim section because a number of the assembly sub-processes was the highest compared to the other two sections.

Table 2: Result of motion time study for material preparation activity in warehouse

Activity	Cycle time (Minutes)
1. Unboxing	5.00
2. Quality Inspection	13.00
3. Segregation	7.00
4. Material arrangement	5.00
5. Transferring process	10.00
Total Cycle Time :	40.00

Currently, as practice by selected case study industry, the Trim section required nine sub-processes, while Chassis and Final Assembly only required six sub-processes for each section. Means, 21 assembly processes are required to assemble a finished-good product. A higher number of sub-processes in production floor meant it required a higher number of raw materials. Based on this supported condition, the Trim section was finalized as a selected

section to implement the developed framework of effective Kanban system.

In Phase 1, motion time study of material preparation activities in the warehouse has been observed using motion time study analysis which are every step of activities was measured and the results show the required cycle time in of material preparation averagely at 40 minutes to complete a set of raw material for a single unit of finished-good product. The cycle time of material preparation was defined based on detailed analysis as shown in Table 2. This information would be used in this phase for the determination of the effective Kanban size.

Once the process cycle time of material preparation in warehouse activity was determined, then the cycle time of production processes was analysed using motion time study analysis in production floor. The motion time study was captured and compiled based on detailed activities for every process was analysed to determine the actual process cycle time with the purpose to calculate the effective Kanban size. The selected production layout was identified consists of 25 processes from five sections to complete a single unit of finished-good product. All related activities and detailed motions practiced by the workers for every process was captured using a stopwatch during the workers performing their tasks and the captured results of the motion time study for the entire processes was converted into average as shown in Table III. As presented in Table III, from the entire 25 processes, the sub-process sequence number 14 was identified as the bottleneck process because it is required 47.25 minutes. This cycle time was considered as the production cycle time because it was identified the highest cycle time compared to others processes. As mentioned earlier, from entire 25 processes, only 21 processes were considered in determination of the effective Kanban size because that processes involved with material replenishment and the other 4 processes (process 22 until sub-process 25) were verification, checking and testing process without involvement of raw material replenishment process and the average process cycle time of these 21 processes is 37.63 minutes. That means, averagely every 37.63 minutes production processes required to replenish with a new set of raw materials from warehouse regularly based on the planned production output.

Table 3: Result of motion time study for selected production layout

Process	1	2	3	4	5
Time (Min)	40.51	22.50	45.20	42.38	32.34
Process	6	7	8	9	10
Time (Min)	36.42	40.16	45.25	42.37	42.20
Process No.	11	12	13	14	15
Time (Min)	47.20	45.00	47.20	47.25	45.56
Process No.	16	17	18	19	20
Time (Min)	25.52	35.00	26.20	25.22	35.00
Process No.	21	22	23	24	25
Time (Min)	47.00	10.00	10.00	8.00	8.00

Successful identification of the production cycle time and warehouse cycle time, the Kanban size was calculated by using Equation (1) as mentioned in the previous section. As derived by Equation (1), the optimum Kanban size was then identified at 0.85 units or can be roundup into 1.0 unit. The results of calculated Kanban size represents material replenishment with minimum Kanban size of 1 unit was required in production floor in order to complete one unit of finished-good product in every process for current production floor as planned output by production floor. The internal transportation activity between warehouse and production floor required raw material trolley as a material handler, as in material replenishment activity, and it known as a material feeder. Based on identified information of Kanban size, then the queuing system theory was conducted in the following phase to analyse the production waiting time during material replenishment processes.

Since the production floor only required a single unit of the material feeder to operating the production processes, the comparison in term of production waiting time was made by increasing number of the material feeder into two units, known as dual feeders. The two units of material feeders were identified based on an assumption of the change-over process which is one of the material feeder was considered as work-in-progress (WIP) during performing an assembly process by production personnel and the other one feeder was parked as standby area for replenishment purpose. Means, each process of production floor was occupied with two material feeders. During the operation of the production system, the emptied material feeder would be transferred into the warehouse for performing the material preparation activity. In the material preparation activity, some processes are required such as unboxing, quality inspection, material segregation; material arrangement and transferring into production floor are required to ensure the requirement was met. At the completion of the first phase, Phase 2 was performed using queuing analysis theory to analyse the waste of production waiting time.

The queuing system analysis was conducted in the selected section of case study analysis based on the determined service rate and arrival rate as identified from process cycle time of production floor and material preparation in the warehouse. The verification process in the determination of the optimum number of the material feeders was simulated based on two conditions based on calculated effective Kanban size from the previous section. Based on the results of Kanban size is 1 unit per feeder, and then the simulation was decided based on two conditions which are a single feeder and double or dual feeders. The results were obtained from this simulation was then analysed by comparison on the average waiting time or idling time of production process due to material replenishment process. In the analysis of queuing theory system, the arrival rate known as λ according to the capacity of the hourly arrival rate of raw materials from the warehouse into production floor. As identified from previous phase, the warehouse only can supply the raw materials from warehouse into production floor at every 40 minutes, and then it was converted into hourly service rate of material replenishment known as μ at 1.5 by using Equation (2). Then, the arrival of raw material from the warehouse was converted into hourly arrival rate known as λ by using Equation (3) resulted as 1.69 based on one hour divided by production average production cycle time which is 37.63 minutes as presented in Table

3 earlier. Then, the optimum number of the material feeder was determined to minimize the wastes of production waiting time or idling time by comparing the idling time between a single feeder and dual feeders as simulated. Based on the results of both conditions, the arrival rate could be different from single and dual feeders because increasing number of material feeder during material replenishment process, simultaneously decreasing the cycle of material replenishment process. This situation required to further analyse of cycle time for material replenishment process and converted into service rate, μ and the results show increasing from 1.5 into 3.0 as using Equation (1). This situation occurred because dual feeders required multiplying by two and the service rate, μ as summarized in Table 4.

Table 4: Result of queuing analysis

	Variables		
Feeder	Service rate, μ	Arrival rate, ʎ	No. of processes, N
Single	1.5	1.69	21
Dual	3.0	1.69	21

The results presented in Table IV then were then further investigated to analyse the wastes of production waiting time by application of Equation (4) until Equation (8) as mentioned in the previous section. Both of the feeders systems were then analysed in details to ensure the justification of the most effective system for material preparation frequency with the purpose of waste elimination and minimizing the production idling for achieving JIT production system and smoothing the production flow. The results of the comparative analysis between both feeders systems were calculated as presented in Table 5.

Table 5: Result of feeder comparison

Parameter	Single feeder	Dual feeders
L	19.11	18.22
W	6.66	4.22

Based on the result obtained from Table V, according to number processes in production floor consisted of 21 processes, the average number of process waiting to be replenished by the material handler is 19.11 processes by using a single feeder. That means that once the raw material was emptied in any process in production floor, the subsequent process needs to wait the next cycle of replenishment process for required to wait completion of replenishment process for another 19.11 processes to fulfill its request. Translating into cycle time, it could be interpreted as a average individual process in production floor waiting to be replenished by a new set of raw material is 6.66 hours per cycle of raw material replenishment process. Simulation of single feeder impacting the production losses time or stoppage time due to the waste of waiting is 6.66 hours per day. Comparing with the simulation results using dual feeders of material replenishment, the average number of production waiting time to be replenished is 18.22 processes, while the average production waiting time to be replenished is 4.22 hours. Both of the simulation results were then compared in term of production lost time or stoppage time. Since dual feeders is only 4.22 hours per day instead of 6.66 hours per day by a single feeder and it was chosen as effective Kanban size based on elimination 36.64% of the lost time or production stoppage. The results of queuing analysis was presented the dual feeders is an optimum number of material feeders to meet the production requirement because of reducing the waiting time at 36.66% in production floor and simultaneously increasing the production capacity. This situation improved the production efficiency in term of capacity planning because closing the gap between actual production outputs versus planned production output.

Once an effective Kanban size and material preparation frequency was determined in the first phase, and number of optimum material feeders was identified in the second phase, then in third phase, the process of standardization was carried out in the production floor by using visual management concept to ensure achieving and sustaining the optimum production efficiency and effectiveness towards JIT system. As resulted from comparison simulation results from single and dual feeders, a number of material feeders was identified as dual feeders because it was recognized as the most effective. Accordingly, the comparison of the number of production output for both situations; single and dual feeders also were analysed. Application of the single feeder, production floor required planning for the stoppage time is every 6.66 hours due the material replenishment activity and the planned production time for this case study industry is 8 hours per day, means the effective production time only 1.34 hours or 80.40 minutes per day which is during material replenishment process, the production was idling. This situation will be impacted on the actual output of production process where the production system can only achieve for 2 units per day based on the effective available time after deducting the lost time divided by bottleneck process as presented in Table III. Comparing by the implementation of the dual feeders for material replenishment process, the effective available production time will be maximized through parallel performing the job tasks during material replenishment process. Thus, the simulation results was recommended the dual feeders must be implementing during material replenishment process in production floor which is one feeder was considered “in-use” and another feeder considered “standby”. The implementation of the dual feeders for material replenishment process in production floor was eliminated the production stoppage time due to waste of production waiting for raw material to be replenished. The effective available production time was increased as planned 3.78 hours or equivalent to 226.80 minutes after deducting 4.22 hours of lost time or stoppage time from planned production time 8 hours. The production output was drastically increased by 150% which is from 2 units into 5 units per day based on calculation from 3.78 hours available hours divided by identified production bottleneck cycle time which is 47.25 minutes as resulted from Table III.

The determined number of optimum material feeders indirectly identified the number of production Kanban size. To ensure synchronize communication between an incoming material warehouse and production floor, the optimum the number of Kanban size must be similar with number of material feeders. Since dual feeders is an optimum solution for material replenishment process, the Kanban size also suggested as 2 units similar to the number of material feeders to ensure synchronize communication and maximizing the production efficiency. According to Kanban size of 2 units per lot and current production capacity only produced 5 units per day based on eight hours

production planning, 1,400 units of finished good products will be achieved per year with planned working days is 280 days. Based on this result, the number of the cycle for part replenishment was determined three cycles per day formulated from the comparison between daily production planning and Kanban size. Three cycles of material replenishment at production floor is mandatory for everyday activity according to dual material feeder concepts and production forecast. The three cycles meant the upstream time cycle of the process must be at every 2.67 hours according to the derivation from available time, 8 hours per day and cycle required, 3 cycles per day.

Based on the observed time, the result showed that the cycle time completion for one cycle of an upstream process according to the seven tasks was 40 minutes, as defined in Phase 1. It meant the material preparation process time per Kanban size or per dual feeders required 80 minutes per cycle upon verified by supervisor and the feeders were then transferred to the production area.

In the process of maintaining the sustainability of the improvement process for material replenishment process, the routing of the material replenishment feeders was designed for easy and quick detection of empty feeder based on designed "triggering system" from the previous phase. The suitable area in the production floor was marked with "ready-in-use" for parking the materials feeder and it was located in the dedicated area known as WIP area. The main purpose of floor marking is for easy triggering by production personnel and ensure on the readiness of the raw material to be used once required. The location of emptied feeder also was designed and located at the next to each process of the production floor and close to the pathway. The identified area in the production floor was marked with "vacant" and using red color marking to trigger the emptied feeder during the replenishing process. Once the emptied feeder triggered by the material handle, then he/she would be communicated with warehouse personnel for the preparing of raw material according to triggering production process. Since the finite frequency of material preparation cycle was applied in production floor, the scheduled time of each cycle for the material preparation was identified based on three cycles of per day as planned by production floor. All the three cycles of replenishment time were defined and fixed on 09:20 am, 12:00 am and 03:40 pm every day to maximize the production output as shown in Table 6.

Table 6: Part preparation scheduling for triggering system and comparison between single and dual feeders

Time	08:00	09:20	12:00	15:40
Replenish Cycle	0	1	2	3

Achieving efficiency and effectiveness of production system required a systematic approach of wastes elimination. In actual industrial practices, most of the industries performed their continuous improvement activities to improve production efficiency by implementing a Kanban system as benchmarked by LM. But, inefficient determine the Kanban size causing miscommunication between the production floor and warehouse which is created the waste of production waiting time. In order to solve this issue, this research was developed and implemented to ensure an effective way in the determination of the production Kanban size will be achieved through calculated effective number of feeders and develop the triggering system of the material replenishment activity. The main focuses of this research is in an automotive production system which is automotive manufacturer required dealing with thousands number of the raw materials from multiple levels of suppliers or vendors. This condition required optimum efficiency and effectiveness of interaction between process to process and suppliers to customer to ensure synchronization of interaction are achieved. Using the developed approach of effective Kanban size, the wastes of production idling causing waiting to be replenished was drastically improved by 36.64% and indirectly optimized material replenishment process and the production output was increased from 2 units into 5 units per day.

Achieving optimum efficiency and effectiveness of production system is the main goal of any management organization in industrial practices as practices through continuous improvement program, but the lesson learned from this research is that the continuous improvement program should not just focus on inside the production floor only, but outside of production floor also must be considered as potential area of improvement activity such as upstream and downstream processes. Organizations are strived to reduce the wastes of production system inside production floor as much as possible. Using this developed triggering system of production Kanban size, the production waiting has been drastically reduced and drives the production floor towards JIT production system. Widely continuous improvement activity focusing on the interruption time of the production process is expected in the long run after implementing the continuous improvement program in all related sections of the production system. As an encouraging result in this research through the development and implementation of the improvement model, the team decides to continue performing more improvement activities project for production floor for other sections in the future research. The positive results from this developed framework in this research, the impacts from obtained results also change the mindset of the industrial practitioners in avoiding "fire-fighting method" of tackling the production issues.

5. Conclusion

In this research article, the continuous improvement activity in the manufacturing industry concentrating on inventory management has been conducted by introducing and implementing the model of effective Kanban size. The developed model was verified and implemented in the selected case study and the obtained results were analysed through utilizing the number of optimum material feeders according to calculated effective Kanban size and analysis of the production idling time using queuing system theory. Two conditions of material replenishment activity were simulated and analysed which are single and dual material feeders in order to achieve JIT production system. The behavior of the production system through communication between warehouse and production floor was investigated in order to ensure the accuracy and reliability of the improvement process. The ultimate objectives are to ensure the implementation of the Kanban system using an effective number of material feeders for production floor achieving optimum efficiency and effectiveness. To ensure sustainability of the developed model the WIP "triggering system" was introduced based on time management and visual management. Successful implementation of the developed framework of an effective

Kanban size based on an effective quantity of feeders and triggering system leads production system eliminating the wastes of waiting by 36.64%. In overall, the results of this research was meeting the research objectives which are effective communication between the incoming material warehouse and production floor was developed through designing the effective Kanban size and the systematic triggering system was introduced based on effective scheduling of material replenishment process.

Acknowledgment

The authors acknowledge the YUTP-FRG grant (0153AA-E36) provided by Yayasan Universiti Teknologi Petronas (UTP), Malaysia for funding the study that resulted in this article. Also highly appreciation extended for selected case study industry and anonymous reviewers for the comments and advises given which lead to the significantly improved the quality of this research article.

References

[1] N. V. K. Jasti and R. Kodali, "Lean production: literature review and trends," *International Journal of Production Research,* vol. 53, pp. 867-885, 2015.
[2] S. Bhasin and P. Burcher, "Lean viewed as a philosophy," *Journal of manufacturing technology management,* vol. 17, pp. 56-72, 2006.
[3] J. K. Liker and J. M. Morgan, "The Toyota way in services: the case of lean product development," *The Academy of Management Perspectives,* vol. 20, pp. 5-20, 2006.
[4] K. Amasaka, ""New JIT": A new management technology principle at Toyota," *International Journal of Production Economics,* vol. 80, pp. 135-144, 2002.
[5] R. R. Fullerton and C. S. McWatters, "The role of performance measures and incentive systems in relation to the degree of JIT implementation," *Accounting, Organizations and Society,* vol. 27, pp. 711-735, 2002.
[6] W. J. Hopp and M. L. Spearman, "To pull or not to pull: what is the question?," *Manufacturing & service operations management,* vol. 6, pp. 133-148, 2004.
[7] B.-R. Lea and H. Min, "Selection of management accounting systems in Just-In-Time and Theory of Constraints-based manufacturing," *International Journal of Production Research,* vol. 41, pp. 2879-2910, 2003.
[8] E. Lander and J. K. Liker, "The Toyota Production System and art: making highly customized and creative products the Toyota way," *International Journal of Production Research,* vol. 45, pp. 3681-3698, 2007.
[9] B. Rouwenhorst, B. Reuter, V. Stockrahm, G.-J. van Houtum, R. Mantel, and W. Zijm, "Warehouse design and control: Framework and literature review," *European Journal of Operational Research,* vol. 122, pp. 515-533, 2000.
[10] Y. Monden, *Toyota production system: an integrated approach to just-in-time*: Productivity Press, 2011.
[11] Z.-L. Chen, "Integrated production and distribution operations," in *Handbook of quantitative supply chain analysis*, ed: Springer, 2004, pp. 711-745.
[12] H. W.-c. Yeung, "Regional development and the competitive dynamics of global production networks: an East Asian perspective," *Regional Studies,* vol. 43, pp. 325-351, 2009.
[13] A. Gunasekaran, C. Patel, and R. E. McGaughey, "A framework for supply chain performance measurement," *International journal of production economics,* vol. 87, pp. 333-347, 2004.
[14] C.-W. Wu and W. L. Pearn, "A variables sampling plan based on C pmk for product acceptance determination," *European Journal of Operational Research,* vol. 184, pp. 549-560, 2008.
[15] K. Dumičić, V. Bahovec, and N. K. Živadinović, "Studying an OC curve of an acceptance sampling plan: a statistical quality control tool," in *7th WSEAS International Conference on Mathematics and Computers in Business and Economics (MCBE'06)*, 2006.
[16] J. C. Aurich, C. Fuchs, and C. Wagenknecht, "Life cycle oriented design of technical Product-Service Systems," *Journal of Cleaner Production,* vol. 14, pp. 1480-1494, 2006.
[17] C.-W. Wu, W. Pearn, and S. Kotz, "An overview of theory and practice on process capability indices for quality assurance," *International journal of production economics,* vol. 117, pp. 338-359, 2009.
[18] M. Rother and J. Shook, *Learning to see: value stream mapping to add value and eliminate muda*: Lean Enterprise Institute, 2003.
[19] J. Knezevic and V. Narayan, "Business performance and maintenance: How are safety, quality, reliability, productivity and maintenance related?," *Journal of Quality in Maintenance Engineering,* vol. 18, pp. 183-195, 2012.
[20] G. A. Zsidisin and L. M. Ellram, "An agency theory investigation of supply risk m anagement," *Journal of supply chain management,* vol. 39, pp. 15-27, 2003.
[21] A. Sinclair and M. Monge, "Concept facility based on single-use systems, Part 2," *BioProcess Int,* vol. 3, 2005.
[22] A. Neely, M. Gregory, and K. Platts, "Performance measurement system design: A literature review and research agenda," *International journal of operations & production management,* vol. 25, pp. 1228-1263, 2005.
[23] M. Rosvall and O. Christensen, "A model for cost effective improvement of a raw material management in the inventory: A Case study," ed, 2012.
[24] R. Nelson, *Probability, stochastic processes, and queueing theory: the mathematics of computer performance modeling*: Springer Science & Business Media, 2013.
[25] O. Handel and A. Borrmann, "Service bottlenecks in pedestrian dynamics," *Transportmetrica A: Transport Science,* pp. 1-14, 2017.
[26] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of queueing theory* vol. 399: John Wiley & Sons, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 307-310 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

A Model of EM Fields from Static Data for Moving Conducting Cylinder

Esmail Mohamed Mohamed Abuhdima[*,1], Robert Prewitt Penno[2]

[1]*Computer Science, Physics and Engineering, Benedict College, Columbia, SC 29204, USA*

[2]*Electrical and Computer Engineering, University of Dayton, Dayton, OH 45469, USA*

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 19 September, 2018
Online: 12 October, 2018

Keywords:
Backscattered field
Conducting cylinder
Rotation and translation

ABSTRACT

The effect of rotation and translation of an electromagnetic (EM) scatterer upon the scattered field is studied for the case of TM polarization. The Franklin and Lorenz transformations are used to analyze the phase and the magnitude of the backscattered field during the movement of the scatterer. Also, a new model is presented that simulates rotation and translation using static data. This stationary data is generated using FEKO. These simulation results of the exact analytical solution are compared with the results of the proposed model using static data. Finally, a comparison between the phase and magnitude of the backscattered field of a scatterer undergoing both translation and rotation for the two different modes (TM and TE) is presented.

1. Introduction

This work is an extension of previous work [1], which studied the behavior of the scattered field from a conducting cylinder in rotation and translation for the case of TE-mode, i.e. the direction of the incident magnetic field is parallel to the axis of the conducting cylinder. However, the current work concentrates on the scattered field of the moving conducting cylinder in the case of TM- mode, where the direction of the incident electric field is parallel to the axis of the conducting cylinder. Furthermore, a comparison between both modes (TE and TM) is presented herein.

It is known that recent research has focused on the study of different scattering phenomena so as to better evaluate the nature of the scattering body. Historically, the Galilean transformation was used to investigated the effect of the rotation of the conducting cylinder by many researchers [2,3,4]. It was seen from measured results how the behavior of scattered field of the smooth metallic cylinder is affected by rotation. This behavior showed that the incident frequency signal is affected by the rotation of the smooth metallic cylinder. According to the measured results, it is useful to have an analytical expression to explain this phenomenon [5]. The previous work found that the pattern of the scattered field of the rotating cylinder is different in comparison with stationary one especially in the case of TE mode [6,7,8]. In this extensional work, the behavior of the scattered field is investigated, when the polarization of the incident wave is TM polarized form. The expression of the scattered electric field in rotation and translation is found using the Franklin transformation and Lorentz transformation. The new model is used to simulate rotation and translation from stationary static data. Finally, the simulation results of this new model and the previously proposed model [1] are compared.

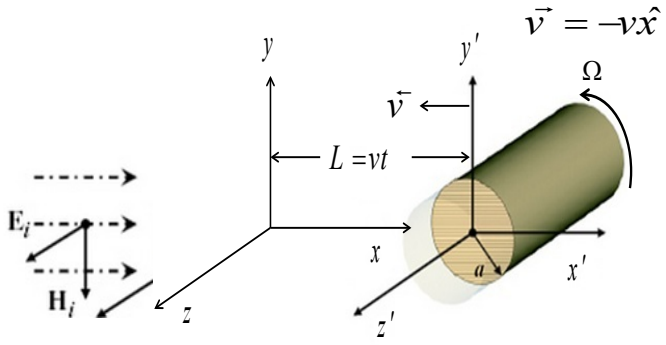


Figure 1: TM -mode of the conducting cylinder in rotation and translation

Figure 1 shows the description of the problem and the incident electric field is represented by [9]

*Corresponding Author: Esmail Mohamed Mohamed Abuhdima, Faculty of Computer Science, Physics and Engineering, Benedict College, Columbia, SC, USA. Email: esmail.abuhdima@benedict.edu

$$E_z^i = e^{-jkx} = \sum_{-\infty}^{+\infty} j^{-n} J_n(kr) e^{jn\varphi}, \quad (1)$$

where $k = \omega/c = 2\pi f/c$ and c is speed of light.

In terms of the Franklin transformation, the scattered electrical field from a rotating conducting cylinder is given by [10]

$$E_z^{sc} = \sum_{-\infty}^{+\infty} B_n H_n^{(2)}(kr) e^{jn\varphi}, \quad (2)$$

where, B_n is a complex scattering number. This unknown constant number is found using boundary conditions at $r = a$, where a is the radius of the conducting cylinder [11]

$$B_n = \frac{j^{-n}\left[c\gamma_n J_n'(\gamma_n a) J_n(ka) - kJ_n(\gamma_n a) J_n'(Ka)\right]}{kJ_n(\gamma_n a) H_n'^{(2)}(Ka) - c\gamma_n J_n'(\gamma_n a) H_n^{(2)}(ka)}, \quad (3)$$

where, $\gamma_n^2 = k^2 N^2 (1 - jq) + \frac{n\omega\Omega}{c^2}(2N^2 - 2 - jN^2 q)$, $N^2 = \varepsilon_r \mu_r$, $q = \frac{\sigma}{\varepsilon\omega}$, Ω is angular velocity, ω is angular frequency, σ is the conductivity, ε is the permittivity, and μ is the permeability. In this case, when the conductivity of the cylinder (σ) is very large, (2) goes back to the original expression of the backscattered field of perfect conducting cylinder in the stationary case as [12]

$$\lim_{\sigma \to large} \mathrm{E}_z^{sc} = E_s^{sc} = \sum_{-\infty}^{+\infty} \frac{J_n(ka)}{H_n^{(2)}(Ka)} H_n^{(2)}(kr) e^{jn\varnothing}. \quad (4)$$

In this step, the approximation to Lorentz transformation is applied to (2) to analyze the phase and magnitude of the backscattered field during translation [13]. The backscattered field of the conducting cylinder in rotation and translation is written as

$$\mathrm{E}_m^{sc} = -\hat{z}\sqrt{\frac{2j}{\pi k'}} \sum_{-\infty}^{+\infty} B_n' e^{jn\pi}, \quad (5)$$

Where,

$$B_n' = \frac{j^{-n}\left[c\gamma_n' J_n'(\gamma_n' a) J_n(k'a) - k' J_n(\gamma_n' a) J_n'(k'a)\right]}{k' J_n(\gamma_n' a) H_n'^{(2)}(k'a) - c\gamma_n' J_n'(\gamma_n' a) H_n^{(2)}(k'a)},$$

$$\gamma_n'^2 = k'^2 N^2 (1 - jq) + \frac{n\omega'\Omega}{c^2}(2N^2 - 2 - jN^2 q),$$

$$\omega' = \omega\left(\frac{c+v}{c-v}\right) \approx \omega\left(1 + 2\frac{v}{c}\right) \quad \text{and} \quad k' = k\left(\frac{c+v}{c-v}\right) \approx k\left(1 + 2\frac{v}{c}\right).$$

The objective of this work is to create a model that simulates translation and rotation of the static data for a complex scatterer as shown in Figure 2. The features of this improved model are written as

$$\mathrm{E}_m^{sc} = \mathrm{E}_s^{sc} \times A(s), \quad (6)$$

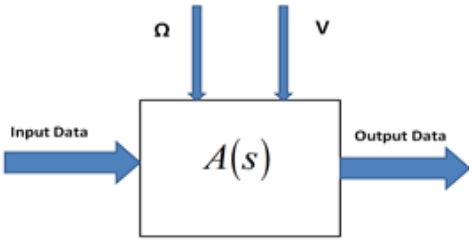


Figure 2: Proposed model

E_s^{sc} is the backscattering electric field of the perfect conducting cylinder in stationary case [12] and $A(s)$ is a transformation given by

$$A(s) = \frac{1}{\sqrt{1 + 2\left(\frac{v}{c}\right)}} \frac{\sum_{-\infty}^{+\infty} B_n' e^{jn\pi}}{\sum_{-\infty}^{+\infty} \frac{J_n(ka)}{H_n^{(2)}(Ka)} e^{jn\pi}}. \quad (7)$$

The definition of the model is given by(7).

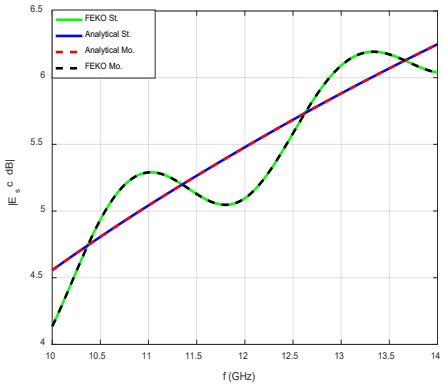
3. Comparison of TM and TE modes

It is interesting to compare results of different types of wave polarization (TM and TE). This paper studies the scattered field of the scatterer under rotation and translation when the polarization of incident wave is in TM-polarized form. The result shows that the backscattered field is changed during the translation of the scatterer with the introduction of the term, $\left(1 + 2\frac{v}{c}\right)$, as shown in (5). Moreover, in the case of the TM-mode, the backscattered field is not affected by the rotation of the scatterer as shown in (5). In the case of TE-mode [1], it was found that the pattern of the phase and magnitude of the backscattered field were affected by the rotation and translation of the scatterer. The effect of rotation shows up with Ω terms as shown in (4) [1], in addition to the effect of the translation as evidenced by the presence of $\left(1 + 2\frac{v}{c}\right)$ terms, similar to the case of the TM-mode. Moreover. the effect of rotation upon the scatterer is more evident for the TE-mode than the TM-mode. Also, the effect of translation of the scatterer upon the phase and magnitude of the backscattered field is evident in both cases (TM and TE).

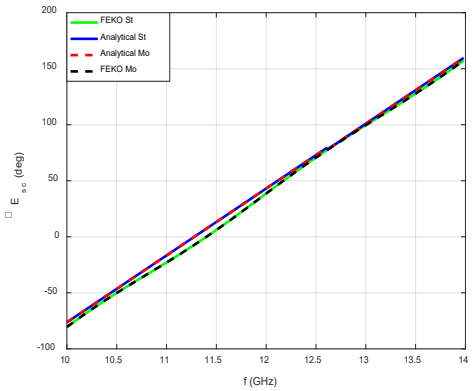
4. Numerical Results

This section includes the simulation results of the new model. The conducting cylinder, which is shown in Figure 1, is modeled using FEKO. It is seen that the incident electric field is z-polarized with the incident plane wave is in the x-direction. It is assumed that

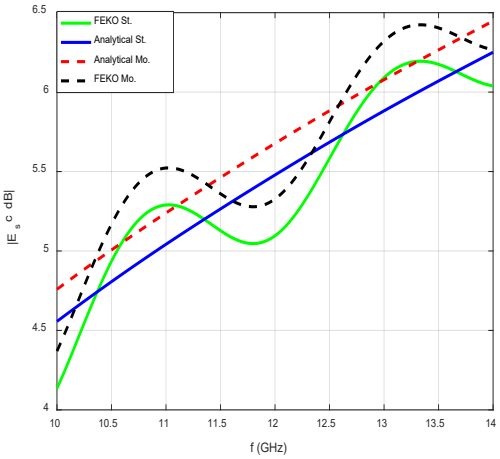
the conductivity of the moving cylinder is $\sigma = 5.76\times10^{7}\ s/m$. The height of the conducting cylinder is $0.1m$ and its radius is $0.025m$. The simulation frequency changed from 10GHz to 14GHz. The static data, generated using FEKO, is inserted into the model as shown in Figure 2 to evaluate the effects of rotation and translation at certain different values of $\beta_a = \Omega a/c$ and v/c, where St and Mo indicate to the stationary and moving cases respectively. Figures 3 and 4 show that the results of both methods are identical and give similar behavior where the slopes of the backscattered magnitudes are equal as shown in Figure 4. Also, the phase and magnitude of the backscattered fields are not affected when the cylinder is rotating with angular velocity $(\beta_a = 0.02)$ and translating with relative velocity $(v/c = 0)$. However, after the relative velocity is increased to $v/c = 0.02$ and $\beta_a = 0.02$, it is seen that the phase and magnitude of the backscattered field are shifted when the direction of translating cylinder is in the positive x-axis as shown in Figures 5 and 6. It is important to see the phase and magnitude when the translation of the rotating conducting cylinder is in the opposite direction. It is assumed that the relative velocity of the rotating cylinder is $v/c = -0.02$ and $\beta_a = 0.02$. It is shown the magnitude of the backscattered field is shifted higher in frequency but it has the same slope as shown in Figure 7. The pattern of the phase is shifted lower as shown in Figure 8.



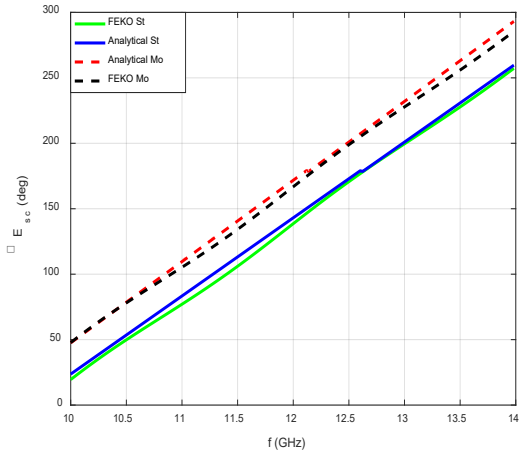
3: Magnitude of TM at $\beta_a = 0.02$ *and* $v/c = 0$



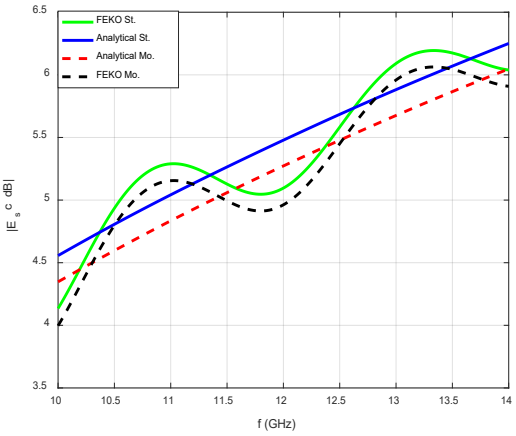
4: Phase of TM at $\beta_a = 0.02$ *and* $v/c = 0$



5: Magnitude of TM at $\beta_a = 0.02$ *and* $v/c = 0.02$

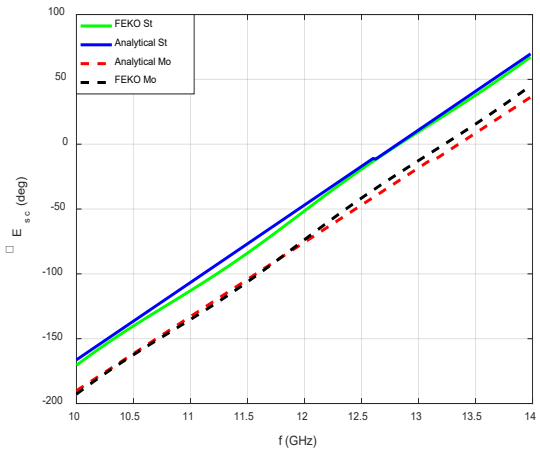


$\beta_a = 0.02$ *and* $v/c = 0.02$

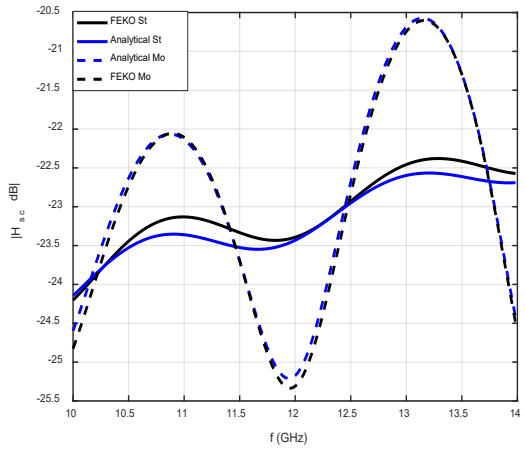


$\beta_a = 0.02$ *and* $v/c = -0.02$

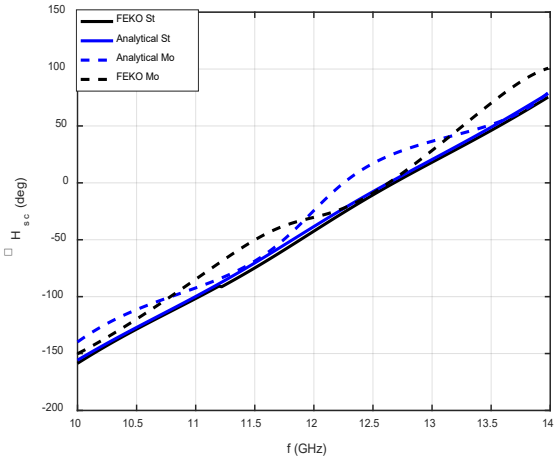
The phase of the backscattered fields for TM mode and TE mode is shown in Figures 6 and 10, respectively. It is noted that the effect of rotation of the conducting cylinder is clearly observed by the periodicity in the backscattered phase in the case of TE mode, the period of which increases or decreases when the speed of the rotation decreases or increases respectively as shown in Figure 11 [1]. The effect of the translation is evident in both modes



$\beta_a = 0.02$ *and* $v/c = -0.02$



$\beta_a = 0.02$ *and* $v/c = 0.02$



$\beta_a = 0.02$ *and* $v/c = 0.02$

(TM and TE) as shown in Figures 5 and 9, and is observed as a shift in the phase and magnitude of the backscattered field. The value of this shift is related to the speed of translation and the direction of the shift is according to the direction of the translation of rotating conducting cylinder. If the translation is along the positive x-axis, the phase is shifted higher as shown in Figures 6 and 10 and the magnitude is shifted lower as shown in Figures 5 and 9.

5. Conclusion

This work studies the effect of the movement of a complex scatterer upon scattering from an incident TM polarized wave. Two types of transformation (Franklin and Lorentz) are used together to analyze the behavior of the phase and magnitude of the backscattered field from the movement of the conducting cylinder. In the TM-mode case, the backscattered phase increases (shifted higher) and the magnitude of the backscattered field decreases (shifted lower) when the direction of translation is parallel to the direction of the incident wave. When the direction of translation is opposite to the direction of the incident wave, the backscattered phase decreases (shifted lower) and the backscattered magnitude increases (shifted higher). The analytical expression and simulation results show that the effect of rotation of the conducting cylinder on the backscattered field is not evident in the case of TM mode. The result of the comparison between the pattern of the backscattered field for both polarizations (TM and TE) shows that the backscattered phase is affected by the rotation of the conducting cylinder, but this effect is more evident in the TE case. The sinusoidal behavior of the backscattered phase is caused by the rotation of the conducting cylinder, whose periodicity reveals important information about the scatterer. Moreover, the magnitude of the backscattered field is affected by the translation in both polarizations. This effect appears as a shift, according to the direction of the translation of a rotating conducting cylinder. Finally, this study attempts to improve the fidelity of modeling the return of an EM scatterer. An extension of this work is planned that would evaluate the effects of the relative movement of the scatterer upon the scattered field in the case of the oblique incident wave.

References

[1] E. Abuhdima and R. Penno,“ A New Model for Simulation of Scattered EM fields from a Conducting cylinder in Rotation and Translation using Static Data,” NAECON-OIS, 2017.

[2] D. De Zutter, "Scattering by a rotating circular cylinder with finite conductivity,"IEEE Transactions on Antenna and Propagation, vol. AP-31, No. 1, PP. 166-169, 1983.

[3] P. Hillion, "Scattering by a rotating circular conducting cylinder I, "Mathematical physics, vol. 41, 1998.

[4] P. Hillion, "Scattering by a rotating circular conducting cylinder II, "Mathematical physics, vol. 41, 1998.

[5] John K. Christensen and Michael J. Underhill,“Doppler Measurements of Smooth and Rough Surface High Frequency Scattering from Spinning Steel Cylinders,” IEEE A&E Systems Magazine , 2004.

[6] E. Abuhdima and R. Penno,“Simulation of the scattered EM fields from a rotating conducting cylinder,” IEEE International Radar, 2015.

[7] E. Abuhdima and R. Penno,“An Improved Model for the Phase of Backscattered Electromagnetic Fields from a Conducting Rotating Cylinder,” NAECON-OIS, 2015.

[8] E. Abuhdima and R. Penno, "Simulation of Scattered EM Field of a Rotating Conducting Number Cylinder using Static Data," Journal of Energy and Power Engineering, Volume 11, Number 6, June 2017.

[9] Harrington, Time Harmonic Elecromagnetic Fields, McGraw-Hill, 1961, pg 234.

[10] Philip Franklin, "The meaning of rotation in the special theory of relativity," Proc. Nat. Acad. Sci., vol. 8, No. 9, 1922.

[11] E. Abuhdima, “Simulation of the Scattered EM field of Moving Dynamic Object using Static Data,” Ph.D Thesis, University of Dayton, 2017.

[12] C. A. Balanis, Advanced Engineering Electromagnetics, John Wiley &Sons, Inc,(1989).

[13] E. Abuhdima and R. Penno,“The Effect of Rotation and Translation upon the Scattered EM Fields of a Conducting Cylinder,” NAECON-OIS, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 308-317 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Synthesis of QDI Combinational Circuits using Null Convention Logic Based on Basic Gates

Duarte Lopes de Oliveira[*,1], Orlando Verducci[1], Vitor Leandro Vieira Torres[1], Robson Moreno[2], Lester de Abreu Faria[1]

[1]*Electronics Engineering Division, Technological Institute of Aeronautics, 12227-000, São José dos Campos - SP, Brazil*

[2]*IESTI, Federal University of Itajubá, 35903-087, Itajubá – MG, Brazil*

ARTICLE INFO

Article history:
Received: 01 July, 2018
Accepted: 04 August, 2018
Online: 12 August, 2018

Keywords:
Asynchronous logic
QDI circuits
Dual-rail code
FPGA
NCL gates

ABSTRACT

Currently, synchronous digital circuits (SDC) may require certain design conditions, such as power consumption, robustness, performance, etc. These design conditions are more difficult to satisfy when SDC are implemented in VLSI (Very Large Scale Integration) technology and in the deep-sub-micron MOS (DSM-MOS) technology. The asynchronous design style has properties that serve as an alternative to design DSM-MOS technology circuits and it can satisfy these design conditions. Quasi Delay-Insensitive (QDI) circuits is a class of asynchronous circuits, they have properties where the DSM-MOS technology design is applied, because they are robust to noise, temperature and voltage variations, as well as low electromagnetic emissions, and they are tolerant to certain faults. An interesting style of QDI combinational circuits are NCL (NULL Convention Logic) circuits, because they accept conventional Boolean functions and it can achieve a better optimization. This paper presents an approach and an architecture based in basic gates for the synthesis of NCL gates, therefore its implementation uses only standard libraries and Field Programmable Gate Array (FPGA). The proposed QDI combinational circuits are implemented in the approach that uses only NCL gates.

1. Introduction

This paper is an extension of work originally presented in *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing* (INTERCON) [1].

Synchronous digital circuits are quite popular and conventionally use a global clock to synchronize their operations. The reason for this popularity is the simplicity of the design. There is also an abundant offer of commercial EDA (Electronic Design Automation) tools for their automatic synthesis. Synchronous circuits have been implemented in deep-sub-micron MOS (DSM-MOS) technology, but global clock signal leads to different problems of synchronous design, such as: a) less tolerance to noise and electromagnetic interaction which increase due to high clock frequency; b) distributing of clock signal is increasingly difficult; c) rise in the delay time variation (Tp-MAX, Tp-MIN) in a gate or wire results in loss of performance, because clock must be adjusted to the maximum delay; d) intensification of drivers amount in the clock signal, consuming a significant part of total energy; e) aggravation of the clock skew; and f) significant delay in wires causes an increase in time analysis complexity [2].

The asynchronous paradigm is a promising alternative to digital design because it eliminates problems related to clock signal and also it increases the robustness of circuits. These circuits operate by events, so there is no overall signal to synchronize operations. Synchronization in asynchronous circuits is performed by protocols of type handshake. In the asynchronous paradigm, there are different classes of asynchronous circuits [3]. The class is defined by its delay model in which the circuit operates correctly and by its operation mode in which it describes how communication works with the environment [4]. The delay models can be classified, such as: a) Bounded Gate and Wire Delay which is similar to the synchronous paradigm [2]; and b) Unbounded Gate and Wire Delay (UGWD). In the UGWD model, the delays of gates and wires can be undefined (any delay value), but finite. In this class, we have the delay-insensitive (DI) circuits that are free of timing analysis. Martin [5,6] shows that DI circuits are very restricted. They can only be synthesized using C elements and inverting gates. Two less restricted variants of the UGWD model

* Duarte Lopes de Oliveira, Email: duarte@ita.br

are: b1) speed-independent (SI) circuits that obey the model in which it has as restriction the delay in wires equal to zero, i.e., the unbounded gate delay and zero wire delay model [3]; and b2) quasi delay-insensitive (QDI) circuits, they have as a constraint the delays in wires when they form a fork, i.e., wires with fan-out > 1 must have approximately equal delays [4,6], so QDI circuits obey the UGWD model and isochronic fork constraint. The interaction of these circuits, for example, DI, SI, QDI with the environment is performed in the I/O_M mode [3, 4, 6]. In I/O_M mode, any change in the output signal can immediately activate a change in the input signal.

For DSM-MOS technology, the asynchronous circuits that best meet the requirements of digital design are the QDI circuits [7]. The QDI circuits have important properties, such as: a) potential to have better latency time, because it works with real delays and not with maximum delays; b) greater robustness to PVT (variations of Process, supply Voltage and Temperature); c) greater robustness to interaction with environment and Stuck-at faults (class of faults tested easily); d) it allows high reuse, because they are highly modular, so they can be used as intellectual property - IP [8]; e) better performance in design of security systems (e.g., encryption) [9]; and f) highly simplified timing analysis.

In QDI combinational circuits (QDI_CC), the data are represented by DI codes of type m-of-n and the most usual processing is the "4-phase" protocol. Several styles were proposed for the synthesis of QDI_CC [10-31]. Firstly, we quote the Minterm Insensitive Delay Synthesis (DIMS) [12] which is quite popular because of its simplicity, but it has a large overhead in area. In DIMS functions, products are implemented with C-elements. Figure 1a shows the operations table of C-element, Figure 1b and 1c show implementations as semi-static CMOS and basic gates, respectively. Another style is the NULL Convention Logic (NCL) proposed by Kant et al. [13]. The NCL style is based on a set of 27 complex gates implemented at the CMOS transistor level [14, 15]. The method for designing QDI_CC based on NCL gates departs from conventional minimized Boolean functions that are transformed into dual-rail Boolean functions and then, the technological mapping is performed using an NCL gates library [16]. The QDI_CC style based on NCL gates has the difficulty of implementing NCL gates in programmable devices, such as Field Programmable Gate Array (FPGAs) [21, 22], while for Very Large Scale Integration (VLSI) it is used standards libraries.

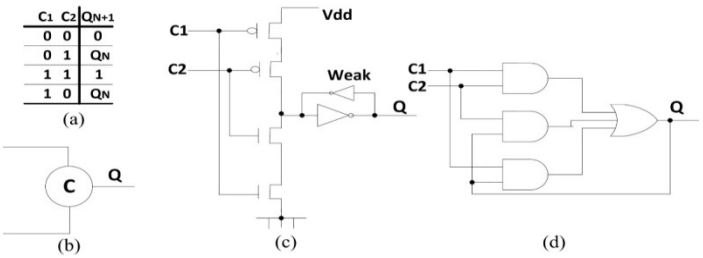


Figure 1. C-element: a) table of operation; b) symbol; c) semi-static version; d) version based on basic gates.

The operating table of NCL gates, not considering the weights, is shown in Figure 2a. The symbol for an THmn NCL gate is shown in Figure 2b, where *n* is the input number and *m* represents the minimum number of inputs that are activated with value one, so that the output is activated with value one. For output to be enabled at zero, all *n* inputs must be enabled at zero. For illustration, the TH23 NCL gate shown in Figure 3 has been implemented in static CMOS technology. In [17-20] three basic gates architectures that implement NCL threshold gates have been proposed for applications that involve standard-cell VLSI or FPGA platforms. In order to operate correctly, these architectures must obey the fundamental mode (FM), therefore they are not QDI.

The interaction with the environment proposes that in the FM mode, "for a new activation of some input signal, the circuit must be stabilized in a stable state, therefore without any electrical activity" [1]. The C-element implemented with basic gates, as shown in Figure 1c, is not QDI because when Q goes from 0 to 1, the circuit must satisfy the fundamental mode to change the inputs again. The QDI_CC design styles of [11, 26-31] are based on basic gates, but they use additional circuits to detect the indicatability which is the requirement to be free of timing analysis.

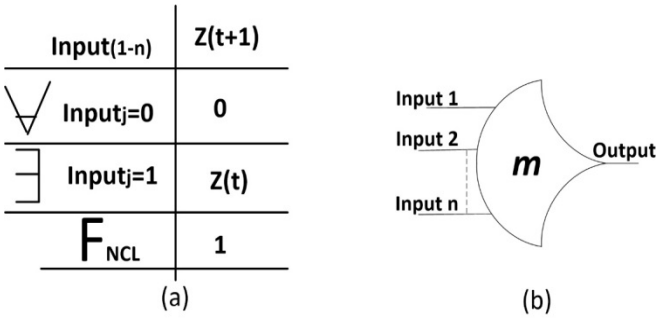


Figure 2. THmn NCL gates: a) table of operations; b) symbol.

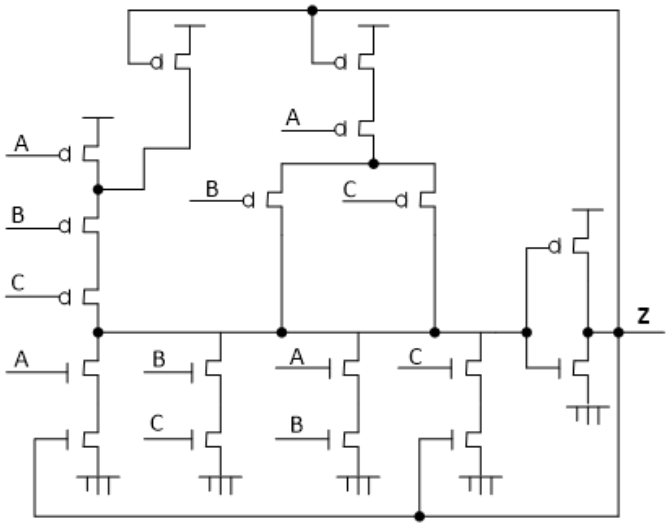


Figure 3. TH23 NCL gate of [16].

To implement THmn NCL gates, this paper proposes a new architecture based on basic gates (see Figure 4). The resulting NCL gates operate in I/O_M mode, therefore they are QDI. Because the NCL gates library is synthesized with only basic gates, then these NCL gates can be easily mapped to platforms of type FPGA or type standard-cell VLSI. This paper also proposes an approach that synthesizes Boolean QDI functions with strong indication using only proposed NCL gates.

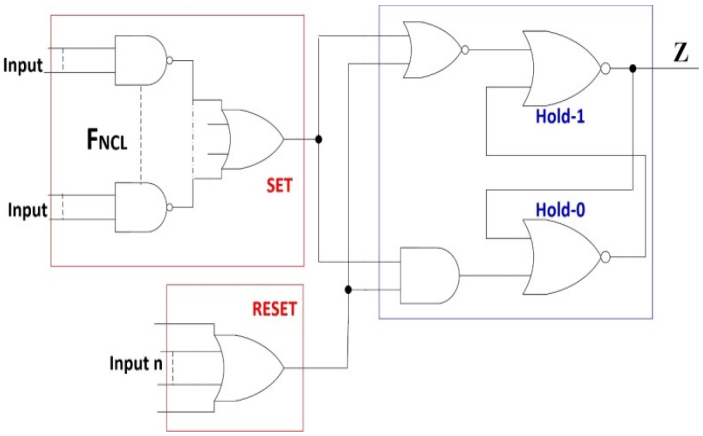


Figure 4. Proposed architecture for THmn NCL gates.

2. Synthesis of NCL Gates: Overview

In order to implement NCL gates, different architectures were proposed at the transistor level [14, 15]. We can also cite architectures that use basic gates, such as proposals [1, 17]. The architecture of [17] is based on Huffman's machine, and the architecture of [1] uses RS latches, which is a semi-static CMOS version of [15].

2.1. Synthesizing Huffman machines as NCL gates

The implemented THmn NCL gates in the Huffman machine architecture [17-20] (see Figure 5) can be synthesized by the Huffman method [3]. We chose the NCL TH23 gate to illustrate method and architecture. The first step of design procedure started with the generation of operations table from the Z function of TH23 NCL gate, as shown in Figure 6a, its function being Z = AB + AC + BC. In the second step, the operations table is represented in the Karnaugh map. In final step, it is realized conventional logic minimization, extracting next-state equations Z(t+1), shown in Figure 6b. The final Z(t+1) equation was extracted by an algebraic manipulation. The logic circuit of TH23 gate is shown in Figure7.

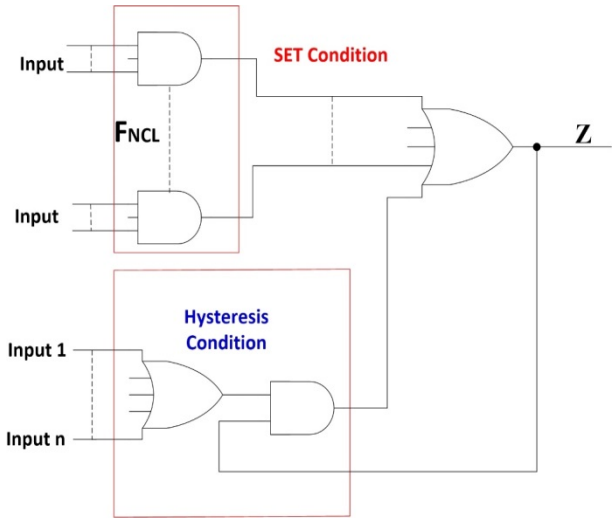
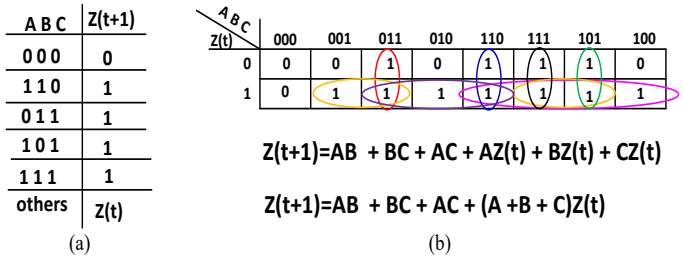


Figure 5. Architecture in the FM for NCL gates of [17-20].



(a) (b)

Figure 6. Synthesis of TH23 gate in architecture of [17-20]: a) Operations table; b) Next-state equation.

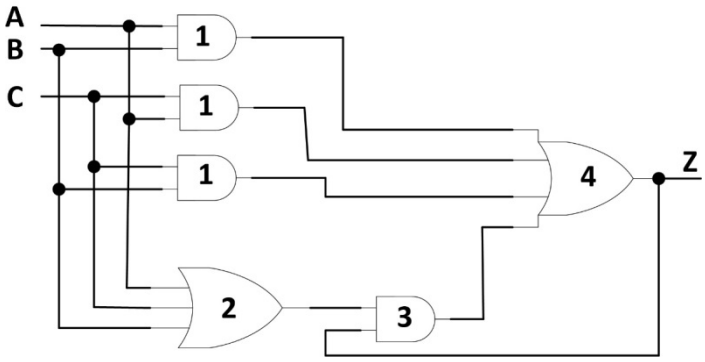


Figure 7. FM architecture of [17-20]: netlist of TH23 gate.

2.2. Synthesizing standard RS as NCL gate

The synthesis of NCL THmn gates in standard RS architecture based on [15] (see Figure 8) extracts the F_{SET} and F_{RESET} functions as shown in [3]. For NCL gate THmn the F_{SET} function is the Z function itself, but using its complemented products. The F_{RESET} function can be described by an OR gate of fan-in=n, which detects when the n input signals are set to zero. The TH23 gate will be used to illustrate method and architecture. Since the Z function of TH23 gate is AB + AC + BC, then the F_{SET} function is {(AB)', (BC)', (BC)'}. Figure 9b shows the F_{SET} function extraction, starting from the operations table and Myers' [3] method, as well as the extraction of F_{RESET} function. The logic circuit of TH23 gate is shown in Figure 10.

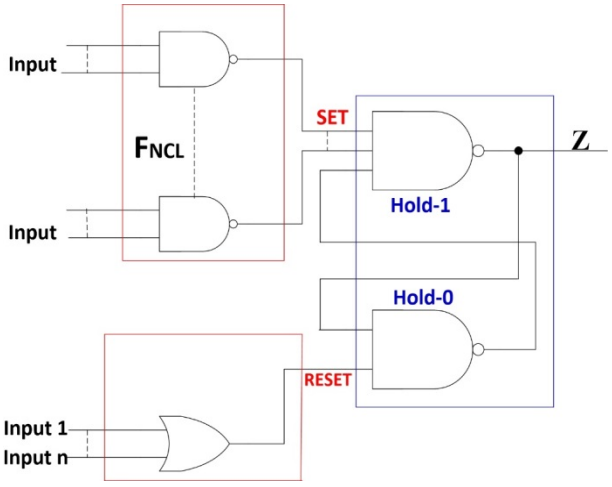


Figure 8. Standard RS architecture in the FM for NCL gates.

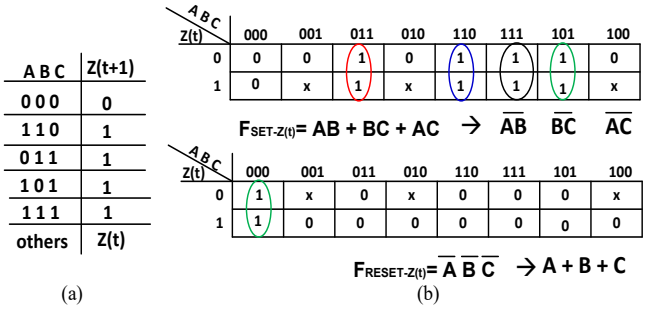


Figure 9. Synthesis of TH23 gate in the standard RS architecture: a) Operations table; b) Fset and Freset equations.

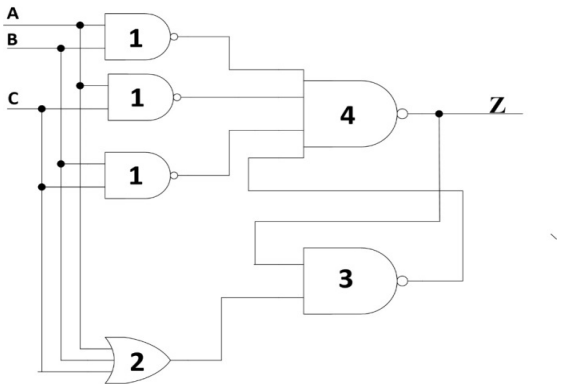


Figure 10. Standard RS architecture: netlist of TH23.

2.3. Synthesizing modified RS as NCL gates

The implementation of THmn NCL gates in modified RS architecture was proposed in [1] (see Figure 11). It follows conventional extraction of the F_{SET} and F_{RESET} functions as shown in section 2.2, but products of F_{SET} are not complemented. To illustrate the method and architecture, it is used the same TH23 gate. Equation 1 describes the architecture of [1] that is composed by the $F_{NCL\text{-}SET}$ and $F_{NCL\text{-}RESET}$ equations. Equations 2 and 3 are extracted using the same procedure described in Figure 9, where $F_{NCL\text{-}RESET}$ is complemented. Finally, equation 4 is obtained from equations 2 and 3, which is the next-state equation of the TH23 gate. Figure 12 shows the logic circuit of TH23 gate in this architecture.

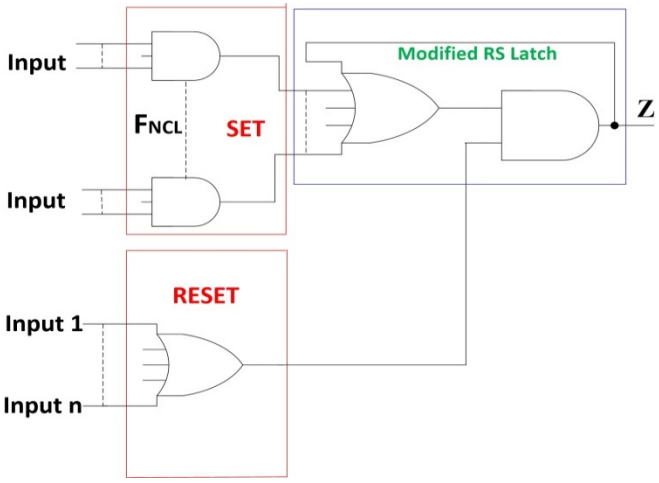


Figure 11. Modified RS architecture in the FM for NCL gates of [1].

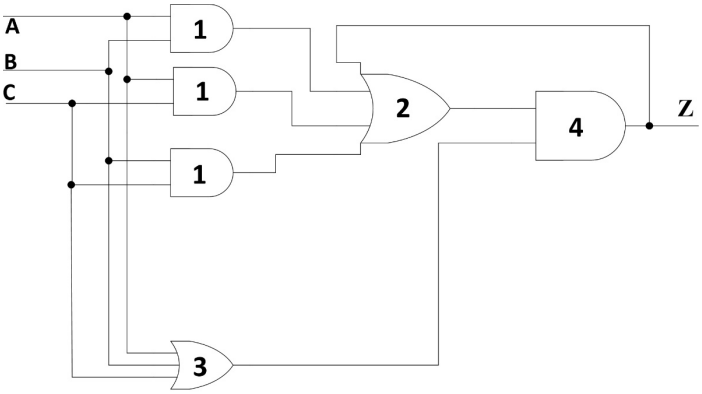
$$Z(t+1)=(F_{NCL\text{-}SET}+Z(t)).F_{NCL\text{-}RESET} \quad (1)$$

Where,

$$F_{NCL\text{-}SET}=AB+BC+AC \quad \text{and} \quad (2)$$

$$F_{NCL\text{-}RESET}=\overline{\overline{A}\ \overline{B}\ \overline{C}}=A+B+C \quad (3)$$

$$Z(t+1)=(AB+BC+AC+Z(t)).(A+B+C) \quad (4)$$



3. QDI Boolean Function: Concepts

Boolean functions for QDI are called function blocks (QDI_FB) and they are synthesized in DI codes. There are different DI codes and in this paper, we adopted the dual-rail coding [32]. The QDI_FB circuits that will be synthesized operate behaving according to the 4-phases handshake protocol [4]. In the dual-rail code, each variable is encoded with two bits. For the variable A, we have A1A0=00 (null - space), A1A0=01 (data 0), A1A0=10 (data 1) and A1A0=11 (never occurs). The DI codes generate the operation completion signal without need of a delay element and with a relatively simple circuit.

3.1. Boolean function: QDI condition

The delay-insensitive (DI) combinational circuits are subject to hazard. Hazardous circuit means that there is a potential for glitches to occur, so it may lead to malfunctioning. The hazard manifests in DI circuits through gate orphan and wire orphan, i.e., a circuit is DI if it is free of gate orphan and wire orphan.

Definition 1 - Gate orphan: Being a function block called FB; it has a gate orphan if a sequence of signal transitions across a path of one or more gates is not recognized by a transition signal on any primary output.

Definition 2 - Wire orphan: Being a function block called FB; it has a wire orphan if the signal transition on a line is not recognized by a transition signal on any primary output.

Since QDI combinatorial circuits satisfy the isochronic fork assumption then these are free of wire orphan, so the combinatorial circuit is QDI if it is free of gate orphan. Figure 13a shows a circuit where input signal transition acts on gates 1 and 2, but it is not recognized by the output signal related to gate 2, so we have a wire orphan. Figure 13b shows the transition of two input signals that activate the OR gate, but it is not recognized by the output signal

related to gate 3, so there is a gate orphan, which means that the circuit is not QDI.

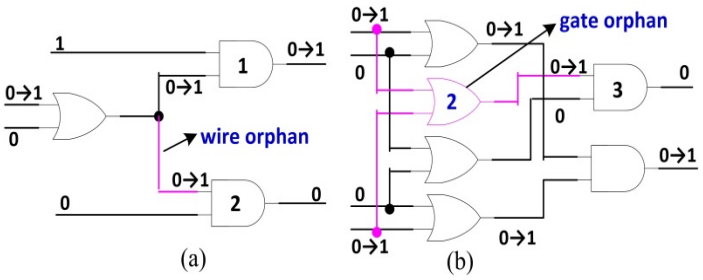


Figure 13. Hazard delay-insensitive of [14]: a) wire orphan; b) gate orphan.

3.2. QDI Boolean function: interaction with environment

A function block for a QDI circuit is required to indicate when its cycles of operations NULL and Valid Data are completed. Hence, by observing the function block outputs, the environment must be able to determine when all input and internal signals have settled. This behavior is shown by a property called indicatability [30, 31]: the signal transition *a* is indicated by *b* since the occurrence of *b* indicates that *a* has already occurred. We say that *a* circuit is indicating if it possesses the indicatability property [30, 31]. The indicatability property ensures that no circuitry outruns a signal which changes slowly. Furthermore, the presence of hazards in an indicating circuit is excluded [30, 31].

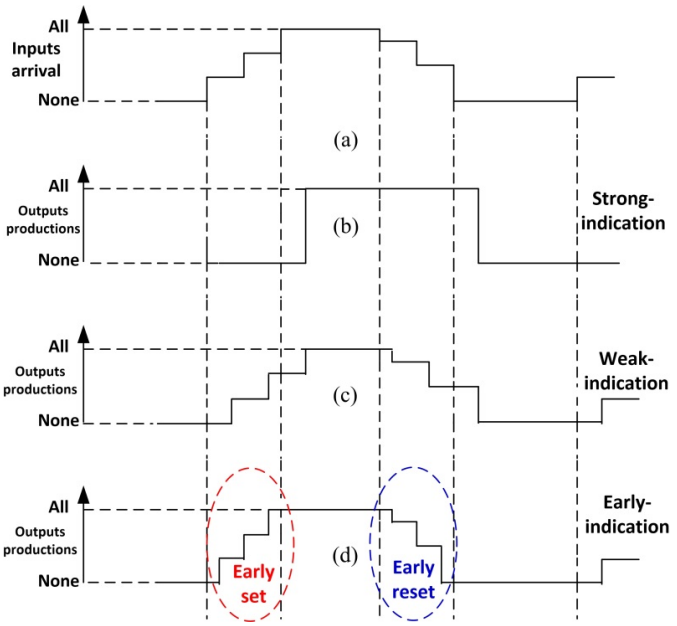


Figure 14. Input-output timing behavior of QDI circuits: a) input arrival; b) strong-indication; c) weak-indication; d) early-indication.

The indicatability property tells you how robust the circuit is in aspect of timing analysis and freedom of orphan. Figure 14 shows the three types of interaction where strong indication is the most robust and, early indication (output) is the least robust. Strong indication means that the output signal transitions will only occur when all input signals are NULL (space processing) or all input signals are valid (data processing). Weak indication means that some transition of the output signal may occur without all input signals being NULL or valid, but for the last transition of the output signal, all input signals are NULL or valid. Early indication means that for all transition of output signals be NULL or valid, the input signals are not necessarily NULL or valid.

If the function block does not indicate inputs on the outputs, it is necessary to add circuitry to ensure correct QDI operation. The inputs are indicated through circuitry with a single Boolean output, whose only purpose is to indicate transitions on the input variables. This circuit is called a status detector. The output from status detector together to the output from function block is used by the environment to ensure correct QDI operation [30, 31].

3.3. Timing analysis

The basic assumption behind QDI circuits is that gates and wires have unbounded unknown delays; the only timing assumption permitted is the isochronic fork – whenever a wire forks to two destinations or more, delays on the forks are approximately equal [3, 4]. In contrast, our approach assumes that the delay of gates and wires are bounded by given time intervals:

($Tp_{MIN\text{-}G}$, $Tp_{MAX\text{-}G}$) for a gate and ($Tp_{MIN\text{-}W}$, $Tp_{MAX\text{-}W}$) for a wire, these intervals represent the lower- and upper-bound delays for propagating an input change to the output in a gate. These delays can either be obtained from standard-cell library characterizations or they can represent the theoretical limits of parametric variation. For simplicity, we despised the delays in wires and we will analyze the three architectures presented in Figures. 5, 8 and 11.

I. NCL gates as Huffman machine

By making a timing analysis of the circuit of Figure 7, we have the inequality 5 that reports QDI constraint of the circuit. Analyzing (5), Tp is the propagation time of a gate, and it can be anywhere in range {minimum, maximum}. If the inequality is not satisfied, we will have glitches in output which occur in reason of the fundamental mode violation. As an example, assuming that A and B go from 0→1 and C=0, Z goes from 0→1 and immediately, A goes from 1→0. So Z can vary 1→0→1→0 if (5) is not satisfied. This case shows clearly that to the TH23 gate to operate correctly it must obey to the fundamental mode, so it is not QDI.

$$2xTp_{MIN\text{-}AND1} + Tp_{MIN\text{-}OR4} > Tp_{MAX\text{-}OR2} + Tp_{MAX\text{-}AND3} \quad (5)$$

II. NCL gates as standard RS

The inequality 6 shows QDI constraint of Figure 10 circuit. By analyzing (6), we can have glitches if the inequality is not satisfied. As shown in the architecture of Figure 5, glitch in the output occurs because of fundamental mode violation, so the TH23 gate is not QDI.

$$2xTp_{MIN\text{-}NAND1}+Tp_{MIN\text{-}NAND4} > Tp_{MAX\text{-}OR2} + Tp_{MAX\text{-}NAND3} \quad (6)$$

III. NCL gates as modified RS

The inequality 7 shows QDI constraint of Figure 12 circuit. By analyzing (7), we can have glitches if the inequality is not satisfied, for the processing of valid data → null data → valid data. As shown in the architecture of Figure 8, the glitch in the output

occurs because of fundamental mode violation, so the TH23 gate is not QDI.

$$2xTp_{MIN\text{-}OR3} + Tp_{MIN\text{-}AND4} > Tp_{MAX\text{-}AND1} + Tp_{MAX\text{-}OR2} \quad (7)$$

4. Synthesis of QDI NCL Gates: Proposed

The design of NCL THmn gates, in case TH23, THand0, TH24comp and TH34w3, will illustrate the approach in the proposed architecture of Figure 4. For THand0 gate, the function is Z = AB + BC + AD while for TH24comp gate, the function is Z = AC + AD + BC + BD. For TH34w3 gate, the function is Z = A + BCD.

4.1. Synthesis of NCL gates

The implementation of proposed THmn NCL gates follows the conventional extraction of F_{SET} and F_{RESET} functions, as shown in section B. F_{SET} function is the NCL gate function itself. F_{RESET} function is the detection of *n* signals going to zero, therefore being an OR gate with fan-in equaling to *n*. Figure 15 shows the logic circuit of NCL TH23 gate where FS_{ET} = AB + AC + BC and F_{RESET} = A+B+C. Figures 16, 17 and 18 show the THand0, TH24comp and TH34w3 gates in the proposed architecture, respectively.

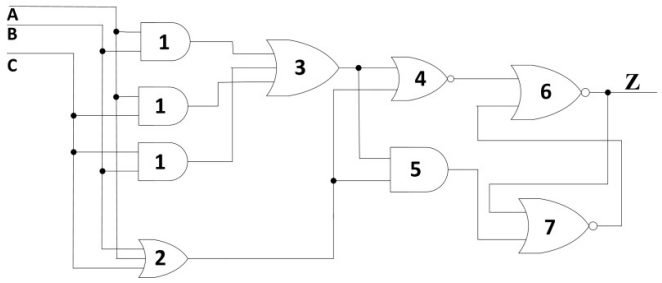


Figure 15. Proposed architecture: netlist of TH23 gate.

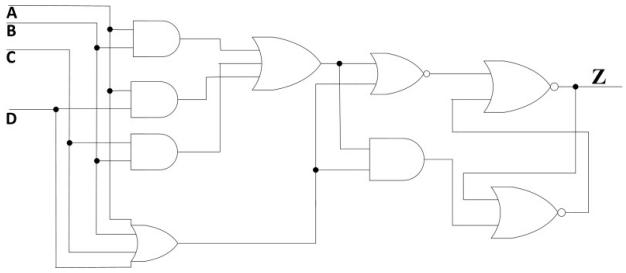


Figure 16. Proposed architecture: THand0 gate.

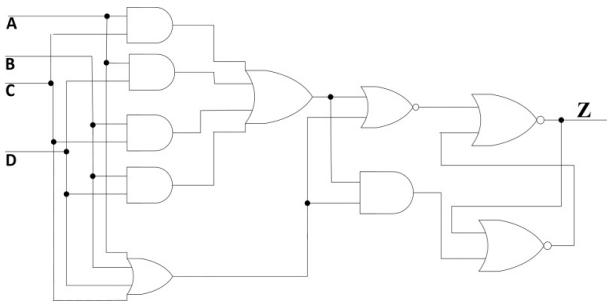


Figure 17. Proposed architecture: TH24comp gate.

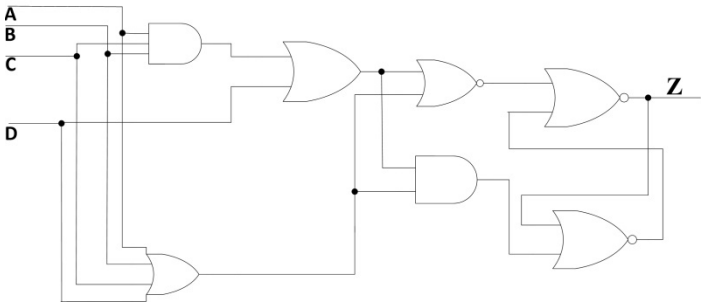


Figure 18. Proposed architecture: TH34w3 gate.

4.2. Timing analysis of proposed NCL gates

Performing the timing analysis in Figure15 circuit, we have that assuming A and B going from 0→1 and C=0, the AND1AB gate goes 0→1, OR3 gate goes 0→1 and NOR4 gate goes 1→0; concurrently the OR2 gate goes 0→1, then AND5 gate goes 0→1 and NOR7 gate 1→0 which the Z output goes 0→1. Immediately, A and B goes from 1→0, so AND1AB goes 1→0, OR3 gate goes 1→0; concurrently OR2 gate goes 1→0 and NOR4 gate goes 0→1 which Z output goes 1→0. For the AND5 gate being QDI, it may be in processing, i.e., still with the value of 1 and consequently, the gate NOR7 with a value of 0, which means that inputs cannot be activated immediately (I/O_M mode). This situation can be measured as shown by inequality 8.

By analyzing (8), we can have glitches if the inequality is not satisfied, for Figure 15 circuit. Since the condition of inequality 8 is robust, i.e., five terms with minimum delay against two maximum delay terms, we can say that it satisfies any DSM-CMOS technology, even if operating at subthreshold voltage, therefore we can accept this gate operating in I/O_M mode as being QDI.

$$2xTp_{MIN\text{-}NOR4} + Tp_{MIN\text{-}NOR6} + Tp_{MIN\text{-}NAND1} + Tp_{MIN\text{-}NOR3} > Tp_{MAX\text{-}AND5} + Tp_{MAX\text{-}NOR7} \quad (8)$$

4.3. Dual-rail NCL Gates

A minimized two-level F function to be implemented as QDI is initially decomposed in basic gates only of fan-in = 2. The basic gates that can be used are: NOT, AND2, OR2, XOR, XNOR, NAND2, NOR2 and AOI4. The seven gates of fan-in = 2 are extended to dual-rail. The dual-rail NOT gate is simply two swapped wires. Each dual-rail basic gate is implemented as NCL gate generating a library of seven dual-rail NCL gates.

The expansion uses the dual-rail code where each signal is encoded in two wires. Figure 19a shows the dual-rail code table for signal a(a1, a0). For a1a0 = 00, it is NULL which means absence of data; for a1a0 = 01 and a1a0 = 10 mean values 0 and 1, respectively. Figure 19b shows the symbol of dual-rail AND4 gate that is implemented from NCL gates knowing that the function F_{AND2} = ab and the inverse function is F'_{AND2} = a'b' + a'b + ab'.

Generating the dual-rail expansion, we have: from F_{AND2}, we obtain $F1_{AND2}$ = a1b1 and from the inverter $F'_{AND}2$, we obtain $F0_{AND2}$ = a0b0 + a0b1 + a1b0. The $F1_{AND2}$ function is mapped to the NCL TH22 gate and the $F0_{AND2}$ function is mapped to NCL THand0 gate. Figure 19c shows the generated THDR-AND2 gate.

Using the same procedure, we can build other THDR-X gates. Figures 20, 21 and 22 show the gates THDR-OR2, THDR-XOR and THDR-AOI4, respectively.

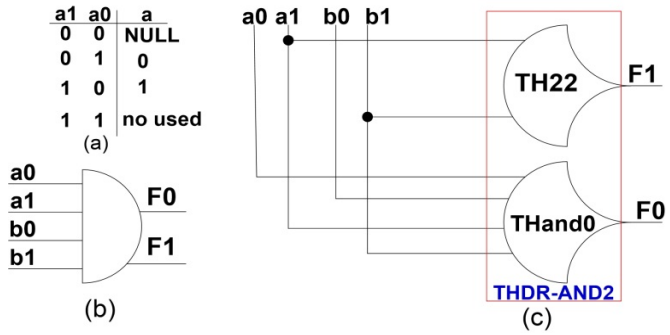


Figure 19. Dual-rail AND Gate: a) Symbol; b) NCL implementation.

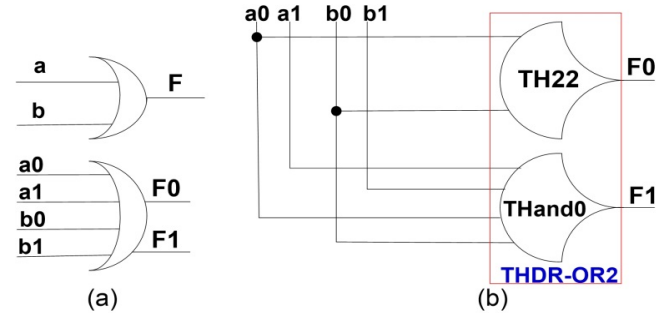
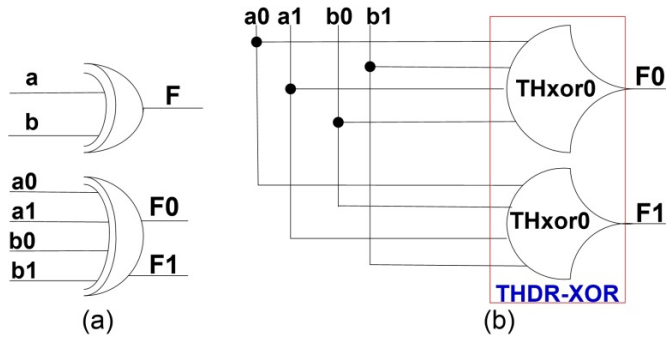


Figure 20. Dual-rail OR gate: a) Symbol; b) NCL implementation.



FFigure 21. Dual-rail XOR gate: a) Symbol; b) NCL implementation.

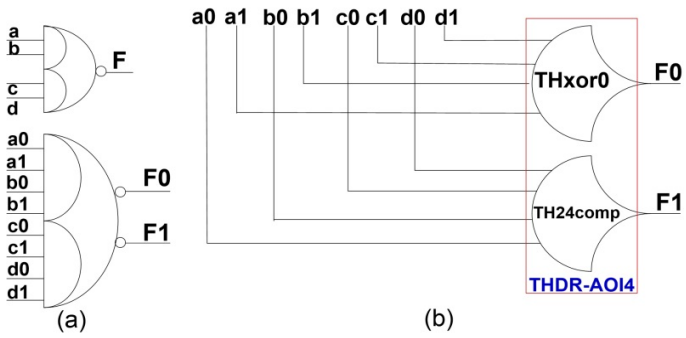


Figure 22. Dual-rail AOI gate: a) Symbol; b) NCL implementation.

Theorem 1. Let the dual-rail NCL gates (THDR-X) synthesized in the proposed approach in section C and implemented in architecture shown in Figure 4. These circuits interact with strong indication.

Proof: Assume all input and output signals are initially in NULL state, since the two functions that make up the THDR-X gate are canonical, then during valid data cycles, every implicant when activated, all input signals acting on the THDR-X must be valid. Since the two TH-X gates contain the reset functions and form the THDR-X gate, then during the NULL cycles all signals that act on the THDR-X gate must go to NULL, therefore the circuit operates with strong indication.

4.4. Proposed Approach for NCL Circuits synthesis

The method starts from a minimized two-level function F_IT (independent of technology) and it follows three steps:

1. Performance of the conventional technology mapping of the $F_{_IT}$ function using only a basic gates library: The mapping is performed, for example, by the SIS [22] tool in the target library [NOT, AND2, OR2, XOR, XNOR, NAND2, NOR2 and AOI4] and to obtain the $F_{_DT}$ function (technology dependent).
2. Performance of the dual-rail extension of each gate of the $F_{_DT}$ function obtaining the $F_{\text{-DT-dual-rail}}$.
3. Performance of the trivial mapping of the $F_{\text{-DT-dual-rail}}$ using a target library of seven dual-rail NCL gates.

To illustrate the approach, we have the minimized function F(A,B,C,D) = A⊕B + CD mapped to a conventional library. The approach performs a mapping that generates for each term of the F function a dual-rail NCL gate and it ends with the NCL dual-rail OR gate. Figure 23 shows mapping to the NCL function and Figure 24 shows mapping of the NCL function to dual-rail NCL. Figure 25 shows implementations in proposed architecture of the NCL gates THxor0, THand0 and TH22, used in the function mapping.

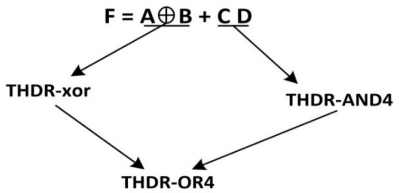


Figure 23. Proposed approach: mapping of NCL gates of F function.

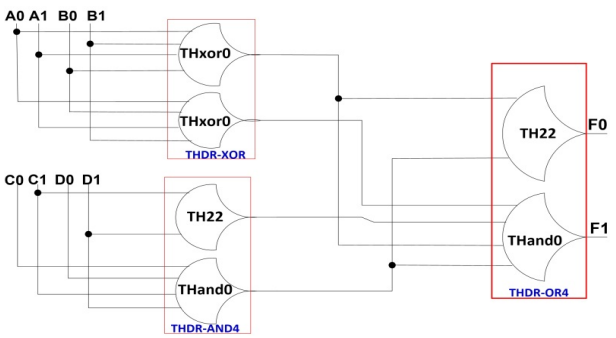


Figure 24. Dual-rail F function: NCL gates.

Theorem 2. Let a dual-rail minimized two-level F function synthesized in the proposed approach of section D with NCL gates THDRmn of section C. This circuit is QDI with constraint of inequality 8 and it interacts with strong indication.

Proof: Assume all the input and output signals of the dual-rail F function are initially in NULL state. The dual-rail F function is free of wire orphan due to isochronic fork assumption. During valid data cycles, the dual-rail F function is free of gate orphan, because of each term of F being associated with a THDR-X gate and there is freedom of gate orphan, with the constraint of inequality 8. As each THDR-X gate operates with strong indication, then the dual-rail function F operates with strong indication, because all terms of the dual-rail F function are activated, either true or false, so all input signals must be valid . For the NULL cycles, the analysis is analogous.

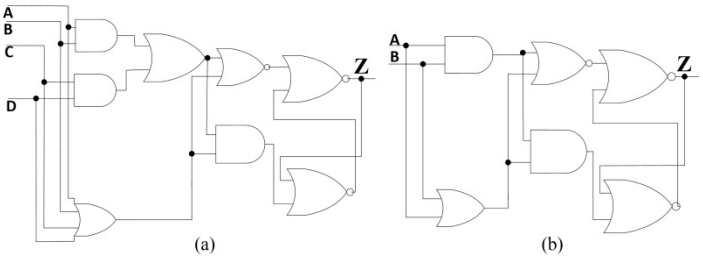


Figure 25. Proposed architectures: a) THxor0 gate; b) TH22 gate.

5. Case Study

In this section we illustrate an approach for the synthesis of NCL circuits. We use an example found in [33] that is a one-bit ALU (Arithmetic Logic Unit) which can be generalized to N bits by using the concept of carry propagation. The ALU is based on the 8-bit ALU of 74181 TTL integrated circuit. Figure 26 shows the table operations with 12 operations being partitioned into three blocks (selection signals M and C0). Figure 27 shows the first step of the approach that is conventional technology mapping generating the multi-level one-bit ALU. The second step transforms the basic gates (single-rail) into dual-rail gates, as shown in Figure 28. The obtained ALU circuit is implemented with XOR, AND2 and OR2 dual-rail gates. The third step performs the trivial mapping of each dual-rail gate into dual-rail NCL gates (THDR-X), as shown in Figure 24, thereby generating the final NCL circuit, as shown in Figure 29.

M=1		M=0	Selection
C_0=0	C_0=1	C_0=X	S_1S_0
A	A plus 1	A	00
$\overline{A}$	$\overline{A}$ plus 1	$\overline{A}$	01
A plus B	A plus B plus 1	A⊕B	10
$\overline{A}$ plus B	$\overline{A}$ plus B plus 1	A⊙B	11

Figure 26. Table of Operations of the ALU presented in [33]

6. Simulation and Results

The design of six benchmarks in two different architectures, in case NCL_D [25], and proposal were synthesized. We also design three NCL gates in four different architectures. These designs were made in structural VHDL, compiled and simulated post-layout in ALTERA tool, Quartus II software, version 9.0, Cyclone III family, in EP3C16F484C6 [34] device.

6.1. Simulations

The simulation post-layout of TH23 NCL gate is shown in Figure 30. Different operations were tested and the waveforms of Figure 30 show its correctness, following the ones shown in operation table of Figure 6a.

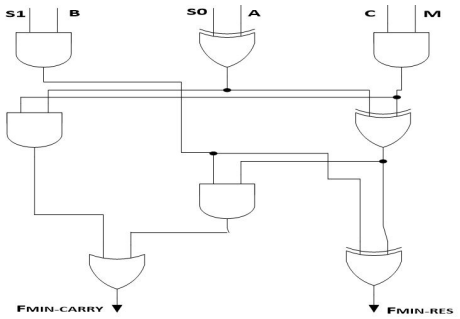


Figure 27. Logic Circuit: Multi-level 1-bit ALU [33].

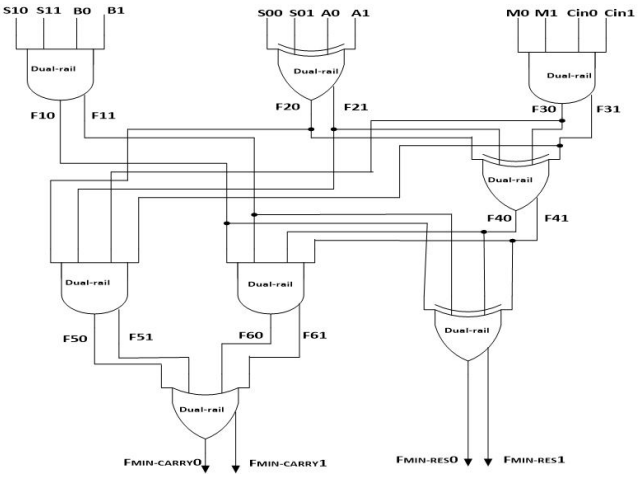


Figure 28. Dual-rail ALU: 1-bit Basic cell.

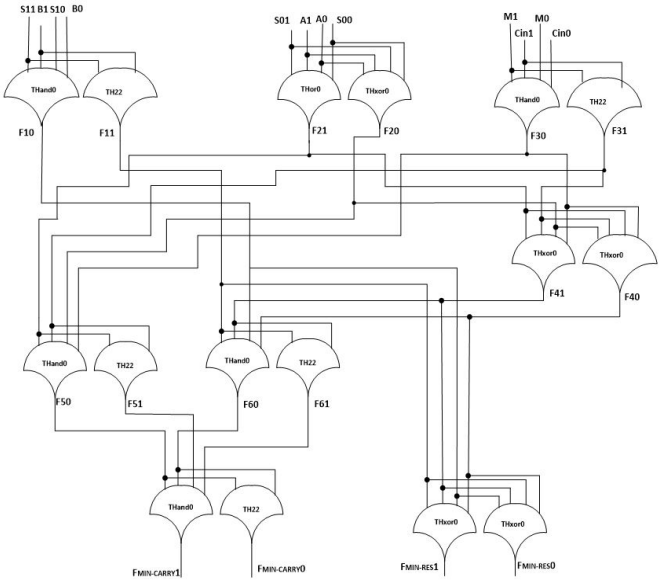


Figure 29. NCL ALU: 1-bit Basic cell.

Figure 31 shows the simulation of NCL ALU. The waveforms show correct results for different tested operations, following the ones shown in operation table of Figure 26. Figure 31 also shows that there are no glitches in the output signal, when compared with the conventional ALU that shows some glitches. The sequence of operations follows the 4-phase handshake protocol, therefore in the timing diagram when all the signals are zero means data absence.

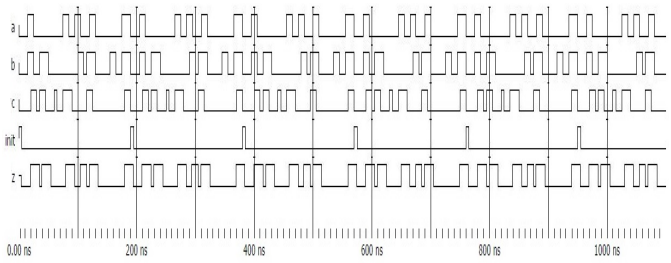


Figure 30. Post-layout simulation: TH23 gate.

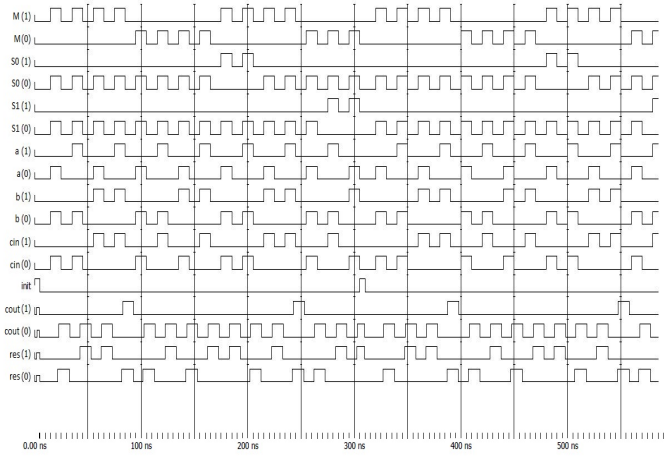


Figure 31. Post-layout simulation: 1 bit ALU.

6.2. Results

Table I shows the results obtained for: area, where only LUTs were used; dynamic power; and latency. Table I is related to the implementation in four architectures of the NCL gate TH23. Analyzing the results of Table I, we have: comparing area in number of LUTs, the proposed architecture, although requiring two LUTs, shows a greater robustness to operate in I/O mode. The proposed architecture obtained an average reduction of 18.6% in latency time when compared with the other three architectures; also obtained an average reduction of 74.0% in dynamic power consumption.

Table II shows the results of six QDI examples design which are: AND of fan-in = 4; detector of prime numbers of 4 bits; detector of odd numbers of 4 bits; full adder of 1 bit; multiplier of 2 bits; ALU of 1 bit of [21]. These examples were synthesized in the NCL_D [25] approach using DIMS components [12] and in the proposed architecture of Figure 4 being both architectures of the type strong indication. The results of six examples in Table II are: number of LUTs; latency time; and the dynamic and static power. Analyzing Table II and comparing the proposal with and NCL_D, we have an average reduction of 13.6% in area (number of LUTs) and 13.7% in static power. There was an average penalty of 4.6% and 32.5% in latency time and dynamic power, respectively.

Table 1. Results of th23 ncl gate in differents architetures

		Time of Latency	Dynamic Power	Number of LUTS
Gate TH23	Figure 7	5.29ns	0.34mw	1
	Figure 10	7.67ns	0.29mw	1
	Figure 12	5.12ns	0.18mw	1
	Figure 15	**4.91ns**	**0.07mw**	2

Table 2. Results of the six examples

	NCL_D architecture				Proposed architecture			
	Nro. LUTs	Latency Time (ns)	Dynamic Power (mw)	Static Power (mw)	Nro. LUTs	Latency Time (ns)	Dynamic Power (mw)	Static Power (mw)
AND4	15	**5.54**	**0.26**	0.051	**12**	5.80	0.36	**0.040**
ODD_4	20	**6.32**	**0.32**	0.067	**16**	7.24	0.45	**0.054**
PRIME_4	65	**7.71**	**1.08**	0.218	**50**	8.38	1.48	**0.168**
ADDER _1 full	27	**6.55**	**0.38**	0.091	**23**	6.63	0.61	**0.077**
ALU_1	43	**8.40**	**1.02**	0.144	**36**	8.83	1.16	**0.121**
MULT_2	95	7.58	**1.49**	0.319	**92**	**7.17**	1.97	**0.309**

7. Conclusions

When comparing the implementation of TH23 NCL gate stated in Table I, our approach clearly shows advantages in terms of latency and power consumption, despite the penalty in area. Therefore, when designs are bounded by achieving the best performance or lower power consumption our approach is advantageous, but for designs bounded by area constraints it is not the best option. Regarding Table II, when it comes to a complete QDI circuit our approach is superior in terms of area.

As a way to increase robustness to FM violation in the NCL gates architectures based in basic gates, it is possible to insert delay elementes in paths that are sensitive to FM mode. This strategy though, has three problems: a) delay elements are not easy to add in FPGAs; b) delay elements degrade the reliability and testability of circuits; c) insertion of delay elements tends to decrease performance and increase power consumption. Thus our approach is more robust because delay elements are not needed.

Due to many different applications requiring robust digital systems in the processing or in data security, an interesting solution shows to be robust components using in their synthesis. In this paper, we present an approach and architecture to synthesize NCL gates on platforms FPGA and standard-cell VLSI. The proposed architecture based on basic gates implements NCL gates that operate in I/O_M mode and they are QDI with restriction of inequality 8 while other architectures based in basic gates generate NCL gates that operate in FM mode, so they are not QDI. Comparing with other approaches [27-31] which need of an additional signal to satisfy the property of indicatability, our approach does not insert any signal. The proposed NCL gates presents other interesting properties such as high robustness to

variations in temperature and to supply voltage, which occurs very often in hostile environments such as in space and in certain areas of military combat. For further works, it is desirable to test the proposed NCL gates considering radiation effects of SEU (Single-Event Upset) in FPGA platform [35, 36].

Conflict of Interest

The authors declare no conflict of interest.

References

[1] D. L. Oliveira, O. Verducci, L. Faria, T. Curtinhas, "A novel null convention logic (NCL) gates architecture based on basic gates" in IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Cusco, Peru, 2017. https://doi.org/10.1109/INTERCON.2017.8079680.

[2] B. H. Calhoun, et al., "Digital Circuit Design Challenges and Opportunities in the Era of Nanoscale CMOS," Proc. of the IEEE, Vol. 96, No. 2, pp. 343-365, February 2008. https://doi.org/10.1109/JPROC.2007.911072.

[3] C. J., Myers, "Asynchronous Circuit Design", Wiley & Sons, Inc., 2004, 2a edition.

[4] P. Beerel, R. Ozdag and M. Ferretti, "A Designer's Guide to Asynchronous VLSI". Cambridge University Press, p. 337, 2010.

[5] J. Martin, "Compiling Communication to Delay-Insensitive VLSI Circuits", Distributed Computing, 1(4), pp.226-234, December 1986. https://doi.org/10.1007/BF01660034.

[6] J. Martin, "The Limitations to Delay Insensitive in Asynchronous Circuits," 6th MIT Conference on Advanced Research in VLSI Processes, pp.263-277, 1990. https://doi.org/10.1007/978-1-4612-4476-9_35.

[7] J. Cortadella, A. Kondratyev, L. Lavagno, and C. Sotiriou, "Coping with the variability of combinational logic delays," ICCD, pages 505–508, 2004. https://doi.org/10.1109/ICCD.2004.1347969.

[8] W. Hardt, et. al., "Architecture Level Optimization for Asynchronous IPs", Proc. 13th Annual IEEE Int. Conf. ASIC/SOC, pp.158-162, 2000. https://doi.org/10.1109/ASIC.2000.880694.

[9] L. Spadavecchia, "A Network-based Asynchronous Architecture for Cryptographic Devices," PhD thesis, University of Edinburgh, 2005.

[10] C. L. Seitz, "System Timing," in Introduction to VLSI Systems, Addison-Wesley, pp. 218-262, 1980.

[11] I. David, R. Ginosar, and M. Yoeli, "An Efficient Implementation of Boolean Functions as Self-Timed Circuits," IEEE Transactions on Computers, Vol. 41, No. 1, pp. 2-10,1992. https://doi.org/10.1109/12.123377.

[12] J. Sparsø, J. Staunstrup. "Delay Insensitive Multi Ring Structures", Integration, the VLSI Journal. v15(13), 1993. https://doi.org/10.1016/0167-9260(93)90035-B.

[13] K. M. Fant and S. A. Brandt. "NULL convention logic: a complete and consistent logic for asynchronous digital circuit synthesis". In International Conference on Application Specific Systems, Architectures and Processors, pp. 261-273, 1996. https://doi.org/10.1109/ASAP.1996.542821.

[14] F. A. Parsan and S. C. Smith "CMOS Implementation of Static Threshold Gates with Hysteresis: A New Approach," In Proceedings of the IFIP/IEEE International Conference on VLSI-SoC, Santa Cruz, CA, USA, 7–10, pp. 41-45, October 2012. https://doi.org/10.1109/VLSI-SoC.2012.7332074.

[15] F. A. Parsan and S. C. Smith, "CMOS Implementation Comparison of NCL Gates," In Proceedings of the IEEE International Midwest Symposium on Circuits and Systems, Boise, ID, USA, 5–8 August pp. 394-397, 2012. https://doi.org/10.1109/MWSCAS.2012.6292040.

[16] S. C. Smith, et al., "Optimization of NULL convention self-timed circuits," INTEGRATION, the VLSI journal 37, pp.135–165, 2004. https://doi.org/10.1016/j.vlsi.2003.12.004.

[17] V. Satagopan, et al., "DFT Techniques and Automation for Asynchronous NULL Conventional Logic Circuits," IEEE Trans. on Very Large-Scale Integration (VLSI), vol. 15, no. 10, pp.115-1159, October 2007. https://doi.org/10.1109/TVLSI.2007.903945.

[18] W. K. Al-Assadi and S. Kakarla, "Testing of Asynchronous NULL Conventional Logic (NCL) Circuits in Synchronous-Based Designs," 22nd IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems, pp.215-222, 2007. https://doi.org/10.1109/DFT.2007.40.

[19] W. K. Al-Assadi and S. Kakarla, "Design for Test of Asynchronous NULL Convention Logic (NCL) Circuits," IEEE International Test Conference, pp.1-8, 2008. https://doi.org/10.1109/TEST.2008.4700611

[20] L. Duc Tran, et al., "Null Convention Logic (NCL) based Asynchronous Design – Fundamentals and Recent Advances," International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom), pp.158-163, 2017. https://doi.org/10.1109/SIGTELCOM.2017.7849815.

[21] S. C. Smith, "Design of an FPGA Logic Element for Implementing Asynchronous NULL Convention Logic Circuits," IEEE Transactions on Very Large-Scale Integration (VLSI) Systems, vol. 15, no. 6, pp. 672-683, June 2007. https://doi.org/10.1109/TVLSI.2007.898726.

[22] M. M. Kim, et al., "Design Techniques for NCL-based Asynchronous Circuits on Commercial FPGA," 17th Euromicro Conference on Digital System Design, pp.451-458, 2014. https://doi.org/10.1109/DSD.2014.85.

[23] T. Chelcea, G. Venkataramani, S. C. Goldstein, "Area Optimizations for Dual-Rail Circuits Using Relative-Timing Analysis," 13th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC'07), pp.1-13, 2007. https://doi.org/10.1109/ASYNC.2007.10.

[24] J. Cheoljoo and S. Nowick. "Technology mapping and cell merger or asynchronous threshold networks," IEEE Trans. on Computer-Aided Design of Integrated Circ. and Systems, Vol. 27(4), pp. 659-672, 2008. https://doi.org/10.1109/TCAD.2007.911339.

[25] M. Ligthart, K. Fant, R. Smith, A. Taubin, A. Kondratyev, "Asynchronous design using commercial HDL synthesis tools", Async'00, pp. 114 – 125, 2000. https://doi.org/10.1109/ASYNC.2000.836983

[26] A. Kondratyev, K. Lwin, "Design of Asynchronous Circuits Using Synchronous CAD Tools", IEEE Design and Test of Computers, vol. 19, no. 4, pp. 107-117, 2002. https://doi.org/10.1109/MDT.2002.1018139.

[27] J. Cortadella, A. Kondratyev, L. Lavagno, and C. Sotiriou, "Coping with the variability of combinational logic delays," ICCD, pages 505–508, 2004. https://doi.org/10.1109/ICCD.2004.1347969.

[28] Zhou, Y., Sokolov, D., Yakovlev, A.: 'Cost-aware synthesis of asynchronous circuits based on partial acknowledgement'. Proc. Int. Conf. Computer-Aided Design, San Jose, USA, November 2006, pp. 158–163. https://doi.org/10.1109/ICCAD.2006.320080.

[29] Fu-Chiung Cheng and Chi Chen, "Can QDI Combinational Circuits be Implemented without C-elements?," IEEE 19th International Symposium on Asynchronous Circuits and Systems, PP.134-141, 2013. https://doi.org/10.1109/ASYNC.2013.13.

[30] P. Balasubramanian, and D.A. Edwards, "Efficient Realization of Strongly Indicating Function Blocks," IEEE Computer Society Annual Symposium on VLSI, ISVLSI '08, pp.429-432, 2008. https://doi.org/10.1109/ISVLSI.2008.103.

[31] P. Balasubramanian, "A Robust Asynchronous Early Output Full Adder," WSEAS Transactions on Circuits and Systems, Issue 7, Volume 10, pp. 221-230, July 2011.

[32] T. Verhoeff, 'Delay-insensitive codes – an overview', Distrib. Comput, 3, (1), pp. 1–8, 1988. https://doi.org/10.1007/BF01788562.

[33] H. Taub, "Circuitos Digitais e Microprocessadores", Portuguese edition, McGraw-Hill, 1982.

[34] Altera Corporation, 2018, www.altera.com.

[35] W. Jang and A. Martin, "SEU-tolerant QDI circuits," in IEEE International Symposium on Asynchronous Circuits and Systems, pp. 156–165, 2005. https://doi.org/10.1109/ASYNC.2005.30.

[36] D. J. Barnhart, et. al. "Radiation Hardening by Design of Asynchronous Logic for Hostile Environments", IEEE Journal of Solid-State Circuits, vol. 44, No. 5, pp.1617-1628, May 2009. https://doi.org/10.1109/JSSC.2009.2017005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 2,309-318 (2018)

www.astesj.com

Special Issue on Advancement in Engineering Technology

ASTESJ

ISSN: 2415-6698

Which User of technology? Perspectivising the UTAUT model by application of the SFL language Pronoun System towards a systems perspective of technology acceptance and use

Cheryl Marie Cordeiro*

Centre for International Business Studies (CIBS), School of Business, Economics and Law, University of Gothenburg, 405 30, Sweden

ARTICLE INFO

Article history:
Received: 24 February, 2018
Accepted: 09 April, 2018
Online: 25 April, 2018

Keywords:
Research methodology
Unified theory of acceptance and use of Technology (UTAUT)
Systemic functional linguistics (SFL)
International business
Technology management
Remote service

ABSTRACT

This study applies systemic functional linguistics (SFL) as complementary framework of analysis of technology acceptance models (TAMs). The purpose is to bridge research methodology language in international business (IB) studies and engineering management science. Currently TAMs and its consolidated version, the Unified Theory of Acceptance and Use of Technology (UTAUT) provides for a typology of one user in one context scenario. The need for the UTAUT model to account for multiple users in multiple work contexts in a single framework of analysis was foregrounded in the study of the workflow processes of a remote services business model of a European founded multinational business enterprise (MBE) with regards to its (i) intra-firm improvements in managing remote services cases, and its (ii) extra-firm selling of life cycle management remote services contracts. The Enterprise has global operations in over 100 countries, of which this study focused on its European operations of improving the quality of remote services for the marine industry. Through an application of SFL unto UTAUT, this study illustrates how multiple users in multiple contexts can be analysed simultaneously, and whose behaviours can be accounted for in a single framework of analysis. The combined SFL UTAUT model addresses the initial statisticity of the UTAUT model, whilst at the same time, expands upon current theoretical perspectives of technology use and acceptance that can be applied in practice.

1. Introduction

Technology acceptance models (TAMs) [1-3] and its consolidated version, the Unified Theory of Acceptance and Use of Technology (UTAUT) [4,5] whilst widely applied to various industry context [6-8], currently provides for a typology of a single user to a single context of use of a specific technology [9-12]. In an era of converging technologies and digital platforms across multiple work spaces, a less static, more sophisticated approach towards the study of technology use and acceptance is needed.

The purpose of this study is to bridge research methodology in language in international business (IB) and engineering management science. It applies the pronoun system found in systemic functional linguistics (SFL) [13,14] from language science as complementary framework of analysis to the UTAUT model. Most technology acceptance studies are quantitative oriented studies. This study contributes to the existing literature by extending the UTAUT model applications by use of SFL, a primarily qualitative analysis approach. The effect of applying the pronoun system in SFL unto UTAUT enables for a simultaneous analysis of multiple users in multiple context of use for a single technology. The need for a multiple user, multiple context perspective in the study of technology use and acceptance arose in the context of studying the workflow processes of a remote services business model of a European founded multinational business enterprise (MBE) with regards to its (i) intra-firm improvements in managing remote services cases and its (ii) extra-firm securing of remote services life cycle management contracts. The Enterprise of study has global operations in over 100 countries, of which this study focuses mostly on its European operations of improving the quality of remote services for the marine industry. The simultaneous understanding of intra- and extra- firm UTAUT processes was necessary for a systems

*Corresponding Author: Cheryl Marie Cordeiro, University of Gothenburg, Sweden. Email: cheryl.cordeiro@handels.gu.se

integral view of the overall Enterprise remote services business model.

A systems perspective [15,16] of the remote services business model for the Enterprise was needed, because if there was a bottleneck of product-services efficiency, the Enterprise should be able to identify the area of inefficiency quickly. In order to do this, a *systemic* systems overview of workflow operations is necessary by top management of the Enterprise. This called for a research practice method that could give a systems integral perspective of UTAUT. But the simultaneous application of the UTAUT model posed a challenge in different contexts in its current form due to the linear staticity of perspective i.e. one UTAUT model for one context, targeted to understand one type of pre-defined User. The aim of this research methods study is to illustrate how this linear staticity can be circumvented by the application of the pronoun system in systemic functional linguistics (SFL) as theory and framework of data analysis. This is done by collecting and analysing data from both intra-firm and extra-firm perspectives given by Enterprise and Enterprise-related respondents for remote services technology use and acceptance.

This paper begins with a literature review of UTAUT and SFL as fields of research, focusing on UTAUT and SFL as methods for research. This study contains two methods sections. The first pertains to the Enterprise remote services marine sector as case example in illustration of how the SFL pronoun system can be applied to the UTAUT model to unfold various User perspectives. The second methods section illustrates how SFL can be combined in complementarity to the UTAUT model, to form an integrated SFL UTAUT model in studying technology acceptance and use. This is followed by a findings and discussion section, where the empirical findings are presented incorporating a discussion on the SFL UTAUT combined model. This paper ends with an outline of study limitations, future research directions and conclusions.

2. Literature Review

2.1. Technology Acceptance Studies

The past forty years have seen scholars design theories and models in understanding of the influencing elements of acceptance and use of technologies. During the 1970s, the Theory of Reasoned Action (TRA) was put forth by Fishbein and Ajzen [1], which explained a person's behavioural tendencies with the aim of predicting changes and interpreting particular personal behaviour. TRA was formulated based on the assumption that behaviour is shaped by intentions that in turn depend on personal attitudes and subjective norms. A decade later, the Theory of Planned Behaviour (TPB), was proposed as an extension of TRA, working on the assumption that behaviours could be controlled by certain parameters in context [2]. In similar timeframe, Davis, Bagozzi and Warshaw [3] proposed the Technology Acceptance Model (TAM) to explain the causal relationships between internal psychological variables such as beliefs, attitudes and behavioural intention and actual information technology (IT) system. The widely studied and considered valid TAM model was based on the two constructs of the User, which was Perceived Usefulness (U) and Perceived Ease of Use (E). These constructs were considered effective when applied to the understanding (even predicting) of individual acceptance behaviour across various information technologies and their users. Subsequent models developed include the combined TAM and TPB that focused on the impact of experience of the use of technology [4], the TAM2/TAM 3 models as a theoretical extension of the TAM that included the perspectives of subjective norms and job relevance when accepting the use of new technologies [5,6], and the Unified Theory of Acceptance and Use of Technology, or UTAUT [13].

UTAUT [6] arose from a comprehensive conceptual review and empirical study of eight technology acceptance models. The unified model consists of six broad constructs deemed to be significant direct determinants of technology acceptance and use that include:

(i) performance expectancy (PE) – the degree to which an individual believes that using the system will help them improve on job performance
(ii) effort expectancy (EE) – the degree of ease of use by the individual of the system
(iii) social influence (SI) – the degree to which the individual perceives it important that others perceive them to use the new system
(iv) behavioural intention (BI) – the degree to which the individual intends to use the system
(v) use behaviour (UB) – degree of affect on the part of the individual when using the system
(vi) facilitating conditions (FC) – the degree to which the individual feels they have the resources and support (technical / organizational) to use the system

An additional four constructs that moderate technology acceptance and use are gender, age, experience and voluntariness of use. Figure 1 illustrates UTAUT as it appears in Venkatesh *et al.* [13:447]. UTAUT continues to be widely used across various technology management studies even if other models of technology acceptance such as the Model of Acceptance with peer support [14] and the Content Acceptance model [15] have been proposed.

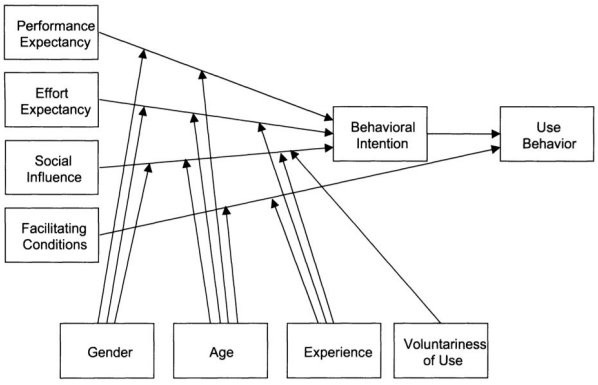


Figure 1. Model of Unified Theory of Acceptance and Use of Technology (UTAUT) [13: 447].

The UTAUT model generally focuses on the causal (cause and effect) relationship between individual attitudes towards using a technology, personal tendencies towards using a technology, actual use of a technology and identifying performance expectancy of a technology. In this model, FCs are taken as the main determinant factor in the use of a technology or system [6].

But while the model of technology acceptance predominantly explain a User's behavioural *expectation* [16] and *intention to use*

[13], *perceived ease of use* [17,18] and *actual use* [6,7] in context (*facilitating conditions*), most studies have implicitly defined the User in a linear model of product-service to the User. In this construct, the User is often defined as the business enterprise end-customer/user that includes a broad range of social actors that include customers [8,19,20], teachers / students [21-24], academics, physicians and nurses [25-27], civilians and military personnel [28,29].

The linear product-service workflow from enterprise to customer as (end-)User disregards within enterprises users who range from service technicians to engineers and product lifecycle managers. This internally defined enterprise users of technology are those who work in support of or are even the creators of the technology for enterprise end-customer/user. The enterprise internal users need to use the same technologies or technological platform to support the product-service workflow from enterprise to end-customer/user. The linear UTAUT concept flow from Engineer to Customer use poses a challenge to the model's inability to overlay contexts of use from intra- to extra- enterprise environments, leaving the perspective of Engineer to Engineer as a field of knowledge under researched.

2.2. Systemic Functional Linguistics

Language is an open adaptive system of choice that humans use language to describe and circumscribe reality [30-32]. We use language in its functional purpose both as process of acting/transacting through time, and as product that helps us situate and identify ourselves in relation to others [33-35]. "Functional linguistic theories are predicated on the claim that language is first and foremost a means for communication between human beings, and that this fact has a deep and all-pervasive influence on the forms that languages take." [36:619]. Language is a resource for making meaning and meaning resides in systemic patterns of options so that language presents itself as a system network potential for meaning [9,10,40].

The usefulness of the UTAUT model when used in research design and method is that it condenses the semiotics (meaning) attributed by Users to technology used in context. Applying UTAUT helps researchers find answers to *why* and *how* Users accept and use technology. User experiences and interactions with technology are expressed through the language system. Because the language system is open and adaptive, the factors abstracted and condensed into UTAUT more often condenses and delimits the full experiential context expressed by Users. In a complementarity, systemic functional linguistics (SFL) is a theory and model of language as a social semiotic system. SFL as theory allows us to account for how language enables us to communicate with each other and with our surrounding in the manner that we do [36]. In that sense, language is its own meta-language, because the language system can be used to study language typology, as well as subjects of other disciplines i.e. other systems. In order to simultaneously apply UTAUT in different contexts of use, this study turns to the underlying unifying theory and framework of language, reflected in the SFL theory and model of language.

The architecture of language reflected in SFL comprises sets of systems and options. When people talk, what is said is usually derived from systemic choice, and SFL gives an open adaptive framework to a general theory of meaning. Language in use is a "system network [that] can represent any domain of activity where choosing can be analysed into small closed sets of options. This does not imply that all such sets of options will be independent of each other – in language they never are." [30:20]. To that end, it would be the consistent dialogic of text and talk carried out by individuals, between individuals and groups of individuals that would also have the power to influence, circumscribe and shape the intersectional relations between producers and users of technology product-services.

Digitalization and the context of Industry 4.0 creates an environment of converging technologies that increases interconnectivity. This increased connectivity between humans and between humans and technology, is also reflected in parallel with the meaning creating system in language when systems simultaneous (rather than dependent) functioning increases its semiotic potential [37]. As such, in the case of the application of UTAUT in understanding technology acceptance, what is needed is a more holistic perspective of not just how a set of defined Users use and accept technology, but how the technology in itself is shaped by feedback from Users in a process of co-evolution of technology product-services. SFL is used in this study to show how it can provide a meta-language framework that allows for the investigation of these dialogic processes between Producers and Users of technology in a context that form today's system of modern ecological habitus. As Bourdieu's [38 81ff] notion of *habitus* has it, "logogenesis provides the material (i.e. semiotic goods) for ontogenesis, which in turn provides the material for phylogenesis; in other words, texts provide the means through which individuals interact to learn the system" and it is through this individual heteroglossic aggregation that a social system evolves. What needs to be investigated in such an interconnected habitus is simultaneity in UTAUT application. This is the ability and capacity for research design and methodology to perspectivise systematically, all actors in all contexts, so that one part of the system can be studied in relation to other parts of the system, larger or smaller. This study focuses in particular, the pronoun system of language that perspectivises deictic points of view such as *I (You), We (They), It* and *Its* applied to the simultaneous application of UTAUT in different contexts. The research questions addressed are:

RQ1: How can SFL be applied to the UTAUT model towards simultaneous use to reflect intra- and extra- firm perspectives of User acceptance and use of technologies?

RQ2: What contributing factors can be established by using SFL as language theory and framework to broaden the applicability of UTAUT into various contexts of use?

3. Method: Case Example

3.1. Remote Services in the Marine Sector as Enterprise Case Example

The business model of the Enterprise of study since its founding in the late 1800s is sales of products from manufacturing in a business to business context. In the past twenty years, its business model has needed to shift from manufactured products alone to digitally connected product-services. The operations of the Enterprise spans over 100 countries with about 135 000 employees working in global teams that speak different languages

across different business sectors. In the era of digitalisation and the Industry Internet of Things (IIoT), the Enterprise began to face differing national and regional data sharing policies as potential barriers to efficiently providing customers with advanced product-service business solutions. Empirical data for this study comes in the form of semi-structured interviews and shadowing of engineers from field studies doing remote services and product maintenance in the marine sector. The field studies were conducted in the Enterprise's office locations for the marine sector. These offices were located in 5 European countries that include Finland, Netherlands, Norway, Sweden and Switzerland (regional remote service centre headquarters location).

Remote services in the Enterprise case example context is defined as a type of system maintenance that allows for a round-the-clock remote monitoring and observation of Enterprise manufactured products. It is part of the Enterprise's advanced services portfolio for the marine sector. Built on big data analytics, such systems of maintenance potentially allow for longer upkeep times of automation processes, sending out alarm signals to Enterprise system engineers, reporting potential system failures and predicting hardware life spans so that components can be changed prior to breakdown. In its efforts towards providing advanced services to its customers, the Enterprise is currently in the phase of building shared digital platforms, both intra- and extra- Enterprise.

The concurrent building of software platforms internal and external to the Enterprise in address of Enterprise efficiency, gave rise to a synchronous duo-User of technology scenario with an overlapping work process timeframe. The first scenario (ScenA) is of a need for a shared internal remote services platform for improved work efficiency for Enterprise employees. ScenA will help employees better coordinate their work efforts across departments, across local business units (LBUs) and its European regional remote services centre located in Switzerland. The second scenario (ScenB), is of a shared external platform for customers that interfaces with ScenA. ScenB is so that the Enterprise remote services team can have continuous contact with its global clients when the vessels are out at sea. ScenA and ScenB Users comprise this duo-User scenario.

3.2. Respondent Profile

In terms of the UTAUT model, what was needed was thus a congruent, synchronous understanding of User acceptance and use of technologies for both ScenA and ScenB, for which the User is differently defined. ScenA would have User defined as Enterprise employees. ScenB would have User defined as Enterprise end-customer, purchasers of the Enterprise remote service systems. Both sets of defined Users for ScenA and ScenB are not homogenous groups of individuals because ScenA for example consists of different employee profiles from Enterprise top managers, system engineers to customer service personnel. ScenA User profiles might also be regionally dispersed, working in different Enterprise business units from headquarters to local business units (LBUs). With most sea faring vessels regionally unconfined, ScenB User profile is also variegated to customers who are located in different parts of the world and who speak different languages. It is not unusual for ScenB Users to encounter different vessel product system use and regional policies. ScenB respondent feedback for this study comes the Enterprise's Product Life Cycle Managers who act as points of contact between the Enterprise and its end-customers. Enterprise end-customers can be located as far as Singapore in Southeast-Asia and are referred to the closest remote services centre in times of distress signals or need of maintenance.

Although the Enterprise has different departments managing remote services that could in total include more than 400 employees in different job functions, there was a challenge in setting up field studies that coincided with respondent time availability for interviews and shadow operations. Enterprise Engineers for example, had to be onsite for both interviews and shadowing of operations to be conducted. As such the total number of respondents available for representation in this study was limited in relation to the given research project timeframe. Interviews and focus group discussions were held with the following Enterprise individuals that categorised as ScenA Users:

- 9 Remote Services Engineer (Headquarters and Local Business Units). These individuals are located at global Enterprise headquarters (Switzerland) that also serves as regional remote service centre for the marine sector. Engineers have different types of specialist knowledge and expertise levels, ranging from Engineering Level 1 (for expert knowledge) and Engineering Level 2 and 3 for lower expertise levels and experience. Cases that cannot be solved for the end-customers are escalated in accordance to specialist knowledge from LBUs to the global centre. Engineers are on call for end-customers 24/7, with no allowance for more than 2 hours downtime for the end-customer. Some engineers are also building the remote diagnostics and maintenance platform whilst the Enterprise is providing this as product-service to the end-customers, so that the engineers in this aspect are both Producer and User of the (same) technology platform.
- 3 Field Engineers. These individuals are located both at global headquarters and at LBUs. Depending on area of expertise and knowledge, the field engineer closest to the vessel is deployed. They are on call 24/7, and can be flown (helicopter or private plane) to vessel sight with immediate notification. Some field engineers in the Enterprise have moved on to the role of Product Life Cycle Managers.
- 2 Marine Remote Service Customer Service (RSCS) Personnel. These individuals are usually the first point of contact between Enterprise end-customers the remote services engineers. They might also receive and manage calls for Enterprise related questions, not pertaining to remote services. Most calls from Enterprise end-customers for remote services in the marine sector are time critical. The remote services team have a response time of maximum two hours to get the parts/components onsite to the vessels. The challenge for these respondents is to get the correct connection between Enterprise expert engineer for its end-customer. They are Users of the Enterprise internal general computer services platform. Calls are almost always initiated by the Enterprise end-customers.
- 2 Enterprise top managers (Headquarters). Top management of the Enterprise consist of a team of individuals in leadership position, although all technological decisions have to be

approved by the Enterprise Chief Technology Officer (CTO) regardless of Enterprise division. The respondents here are important as decision makers and in their capacity to steer technology strategies for the Enterprise for long-term sustainable business. Their main interest is to develop a future internal standardised Enterprise interface that connects the departments of various divisions (even beyond the marine sector). This internal interface should have a corresponding standardized external interface for when end-customers login to their accounts. As such, some aspects of remote services need to have a consistent Enterprise branding and theme that is both employee and end-customer user friendly. A consistent brand and theme for digital interfaces for the Enterprise has been challenging to achieve due to the relative autonomity of LBU workings, and different workgroups within the Enterprise.

Enterprise respondents categorised for ScenB Users include:

- 6 Enterprise Product Life Cycle Managers (LCM). These individuals often have advanced engineering degrees and some have worked as field engineers prior to becoming LCMs. Their role is different from the RSCS personnel in the sense that the LCMs take care of end-customer needs from product purchase to product life-end or recycle of components. Part of their job is to sell diagnostics and service maintenance contracts. Acting as key account managers, they are in long-term direct feedback contact with Enterprise marine sector end-customers (EeC). Their feedback for this study is assumed reliable with regards to feedback from end-customers for the following reasons – (i) LCMs are in constant contact with EeCs due to an earnest effort in improving product-services portfolio for future technology developments, and towards future Enterprise sustainable business competitiveness and (ii) LCMs are held responsible by EeCs for product-service downtime, a poor follow-up of which might affect future service contract agreements.

Respondents were targeted specifically for product-service expert knowledge and because of experience and exposure to both intra- and extra- systems use that was to be improved upon by the Enterprise. All respondents are involved in providing advanced services in the Enterprise in different capacities and areas of expertise, the number of years spent with the Enterprise ranged from 2 to 24 years. Those who have spent 2 years at the Enterprise had mostly joined as master thesis students, who had then gained expert knowledge in a specific field or product before being fully employed by the equal opportunities employer Enterprise.

4. Method: SFL Pronoun System Combined With Constructs of UTAUT

The pronoun system of language in use, which is *I (You), We (They), It* and *Its,* serves as a reference point system to indicate point of view in answer to the questions of *who* is acting / saying and *what* about, under *which* circumstance. It is a language referencing perspectivity system that can be mapped in a four quadrant model (Figure 2). All pronoun perspectives in their various forms, singular/plural, subjective/intersubjective, objective/interobjective, can encompass an "inner" and "outer" experience and can be expressed as such. This model, reflecting of the evolving nature of language, is inherently adaptive and relative in perspective, depending upon researcher definition of the unit of analysis of what is singular/plural, subjective/objective etc. Consistent inquiries from each perspective will render a specific type of knowledge that can be classified under the eight primordial methodological perspectives such as phenomenology (singular subjective), hermeneutics, ethnomethodology, autopoiesis, empiricism and systems theory (plural interobjective). As such, the four quadrants can be said to reflect a type of knowledge related to that perspective of inquiry, reflecting both inner and outer worldviews.

Figure 2 maps the different perspectives. Within the SFL system, the semiosis and expression of inner and outer human experiences are reflected through transitivity processes in the metafunction of language, amongst them are acts of doing (material processes), acts of saying (verbal processes), or thinking (mental processes), indicated through use of verbs. Each transitivity processes encompasses agents who act within an activity context and circumstance.

In the Upper Left (UL) quadrant is the singular subjective perspective of 'I'. When applying UTAUT to technology, the assumed User would have this Agency / Actorship of 'I' reflected in the UL quadrant, such as "I use this technology for *x* purposes" or "Using this technology helps me accomplish <task *x*>". Collected knowledge of UTAUT constructs related to User's expectancy of use can be reflected in the UL quadrant. But the perspective is relative, depending on who is the defined 'I'. In ScenA, the 'I' User would be the Enterprise engineers and remote customer services personnel. In ScenB, it would be the Enterprise's end-customer who uses the remote services produced by the Enterprise. The Lower Left (LL) quadrant reflects the plural intersubjective perspective of "We". This collective perspective tends to reflect an ideology of shared perspectives attributed to the proximity of working with others in the same environment, given allowance for slight variation of these shared experiences. Knowledge collected on UTAUT constructs pertaining to social influence (corporate culture) or voluntariness of use (intra- inter-group workings and social norms) are reflected in the LL quadrant.

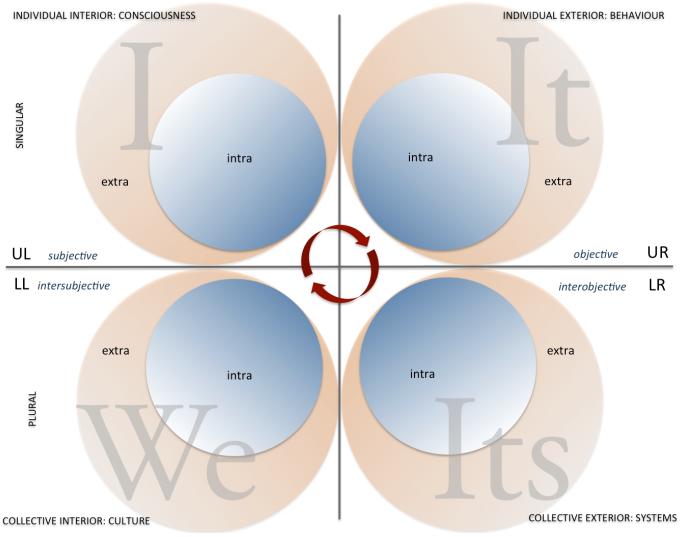


Figure 2. The pronoun system of language in use, mapped in four quadrants (adapted from Wilber [39]).

The Upper Right (UR) and Lower Right (LR) quadrants reflect singular objective and plural interobjective perspectives. Knowledge collected on UTUAT constructs pertaining to a particular technology 'It' and the technology system, 'Its', to which a particular technology belongs are reflected in these quadrants. In this study, it is noted that ScenA presents a perspective from the Enterprise engineers who design and implement the technology and technology system, to which they are Users of these systems themselves. ScenB presents a perspective of the Enterprise end-customer as User who purchases the technology and technology system from the Enterprise. Their technology acceptance experience is different, but both types of Users will need to feedback their experiences of use of technology into the Enterprise system in order for improvement in efficiency of Enterprise remote services delivery.

The perspectives, beginning in the UL quadrant working anti-clockwise presents perspectives that increase in complexity. The perspectives move from *Individual* to *Group* to *Object/s* in a system to a *Systems* view reflected in the LR quadrant. The LR quadrant that reflects a systems view of the subject of inquiry presents the broadest perspective, one that encompasses all other perspectives. All perspectives are in dialogic relation with each other, where a change in one perspective could well influence the context of the other perspectives. The broadest perspective reflected in the LR quadrant encompasses all other perspectives.

The SFL pronoun system model is useful for the simultaneous application of UTAUT in different contexts of use because the four quadrants unfold in a systemic manner, not just the Agency of a material/verbal action but it uncovers (human) Agency in relation to the object of use (UTAUT's type of technology) and circumstance of use (UTAUT's facilitating conditions).

5. Findings and Discussion

Figure 3 shows constructs of UTAUT for ScenA mapped into the SFL pronoun system quadrant, where Enterprise engineers as defined Users. Figure 4 shows constructs of UTAUT for ScenB mapped into the SFL pronoun system quadrant, where Enterprise end-customers are defined Users. The insights for ScenB Users come from Enterprise Product Life Cycle Managers who are in direct contact with the customers. All elements found in the quadrant perspectives are interrelated and in dialogic relationship to each other. The dialogic relations between the Actors are signalled by arrows drawn in circularity. All actions take place through the context of spacetime. Time is indicated as a fundamental background feature in which Enterprise activities and processes take place.

5.1. ScenA UTAUT SFL Pronoun System Findings

Due to that this study is specific in its Users for ScenA being Enterprise top managers, engineers and remote services customer service personnel, the normative definition of User often studied by application of the UTAUT model is slightly modified to reflect the different roles as Users of a single platform system. The single platform system is one that encompasses the marine sector remote diagnostics and product maintenance system both built and used by Enterprise engineers and LCMs. What is different for the different Users of this system is the Time perspective. Enterprise top managers have a longer timeline perspective, in view that one Enterprise technology strategy is to build a standard computer software platform for internal and external use. This standard platform of computer supported services in remote services necessarily interfaces with those used by the Enterprise end-customers, for them to independently login and check on product diagnostics. The Enterprise engineers have different roles, some have project timelines as short as 3 months to come up with computer support. A binding factor for ScenA Users is shared organizational values in technological innovation and service excellence towards end-customers.

ScenA Users hold various expert capacities who worked with the building of the Enterprise products and services that were to be sold to ScenB Users. As such, UTAUT moderating constructs [13] such as 'social influence', 'gender', 'age' and 'voluntariness of use' could have potentially played secondary influencing roles in this study to other factors such as 'experience' that ranked high on the Users list. Most Users for ScenA were expert knowledge workers and the 'social influence' could be redefined as the Enterprise culture of technological innovation, and pioneering remote services work not just in the marine sector but in other heavy industries in which the Enterprise has business operations. UTAUT models also predict User behaviour based on cognitive state of the user, their expectation and intention. These factors for ScenA users were subsumed under a broader corporate culture context of quality excellent in the product-services produced towards end-customers. This then needed a shift in perspective for the application of the UTAUT model towards a broader unit of analysis than the individual as User but towards the Enterprise top management defined as 'User' in this context.

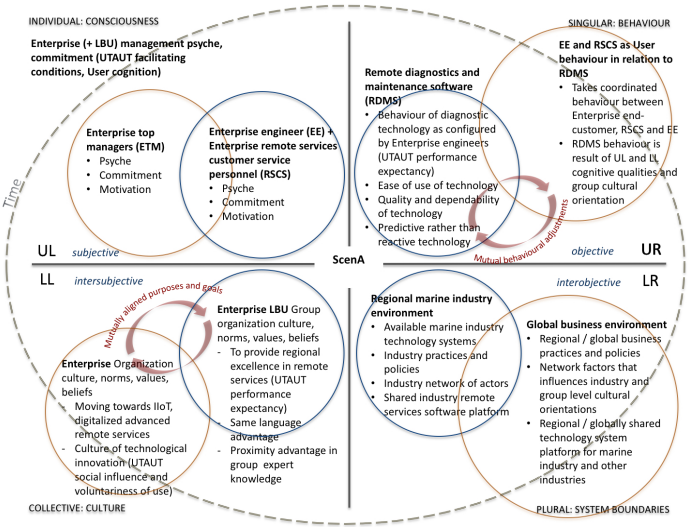


Figure 3. SFL pronoun system perspective of UTAUT Users for ScenA: Users from the Enterprise

The Enterprise top management defined as 'User' towards new technologies comes in the form of Enterprise psyche, commitment and motivation (Figure 3, UL quadrant) of first producing the technology and then using it themselves, as platform towards an integrated advanced services portfolio in an era of digitalisation. To some extent Users of technology will need to be convinced of their intention to use that technology even prior to first testing. In the case of ScenA Users, the respondent feedback from field studies and interviews indicated that the Enterprise engineers

would not (and could not) produce the remote diagnostics and maintenance platform if top management at the Enterprise were not (financially) commitment and believed that this technology would contribute to their competitiveness and to their end-customer's industry competitiveness. It is here that the SFL pronoun system is effective in helping researchers perspectivise and define 'which User' of technology, giving the research model more fluidity and adaptiveness for research purposes, so that the psyche and commitment of the top management of the Enterprise can be accounted for in understanding technology acceptance and use from the Enterprise perspective.

The Enterprise top management commitment reflected in the UL quadrant in Figure 3 not only bolsters both employee motivation and commitment at individual level but it encourages the building of a corporate culture towards constant technology innovation, reflected in the LL quadrant in Figure 3. Enterprise top management discourse would filter through the organization through management hierarchies targeted at group level meetings where Enterprise engineers and remote service customer service personnel would be motivated to provide excellent remote services for their end-customers. This organizational culture of technological innovation creates an understanding towards the building of an Enterprise internal common remote services platform in order to connect LBU employees with its global and regional centres for remote services. The main challenge experienced here, reflected in the LL quadrant in Figure 3, is the communication between the LBUs located in different European countries, and the Enterprise remote services headquarters in Switzerland. Two main challenges arose on why it was difficult building a common internal remote services platform that include (i) preference for local language and (ii) existing remote services technology platform that did not correlate with the architecture of the updated platform as suggested by headquarters.

5.2. ScenB UTAUT SFL Pronoun System Findings

Feedback from Enterprise end-customers (EeC Users, Figure 4) come from Enterprise LCMs who have the role of key account managers, who secure and follow-up on remote services life cycle management contracts with end-customers. LCM interview responses are assumed earnest due to that the proper managing of accurate end-customer feedback is crucial in building future remote services support for both Enterprise and end-customer use. The long-term Enterprise technology strategy is to have a shared platform of remote services between Enterprise and its end-customers. A challenging task much due to that currently, most end-customers are geographic proximity bound by registration of country of ownership of vessel, even if the vessels are globally seafaring. Different types of software systems are used, how much data information the systems can share between themselves are regulated by industry, national and regional policies. The challenge for seafaring vessels and remote services is that vessels lack consistent internet connections due to inconsistent satellite connections. As such, for varying reasons, Enterprise end-customers tend to be contextually (geographic proximity) bound in their immediate business networks when it comes to remote diagnostics and maintenance. Should a sea vessel come into crisis, it is referred to the nearest regional remote services centre or an LBU. Enterprise LBUs, although belonging to the Enterprise, work sometimes in competition with other to secure end-customers. LBUs have the advantage of local language, proximity to end-customers and shared remote services platform systems with their end-customers, built from more than twenty years ago. The strength of the LBU's system is also the weakness of the Enterprise's system due to that LBUs tend to have separate tools and systems to track cases that are not shared with the Enterprise global headquarters. This contributes to the challenges faced by ScenA Users whose current technology strategy is to create a common working platform for Enterprise and its EeCs. While there are Enterprise European initiatives to coordinate between departments and LBUs located in Europe to create a single platform for services, at the time of study, this has yet to be achieved. As such, Enterprise end-customer feedback is crucial and timely information is needed by the LCM in order to work towards a common shared platform for remote services.

The LCM/EeC as User perspective is reflected in Figure 4. The interdependent relations between Enterprise and EeCs are reflected in the UL and LL quadrants in Figure 4, where both Users of remote services need to share a sense of commitment to using the system bolstered by a general movement towards digitalisation and the Industry Internet of Things (IIoT). EeC corporate culture will also need to be one that is progressive and welcoming to technology innovation and change. Some motivating indicators as to why EeC would want to use Enterprise provided remote diagnostic and maintenance service and why they would want a signed contract would be SFL UTAUT constructs pertaining to the UR quadrant of singular objective 'It'. Reflected also in the summary findings in Table 1, what EeC Users find critically important is that the remote services provided is quick (time critical), reliable, user friendly systems that require little effort in learning from the EeC User. This is not to say that EeCs do not wish to put in learning time for acquired Enterprise systems. Rather, that the learning time should be short. EeCs also have as corporate goal, to be independent users of acquired systems, so that they can fix problems themselves without needing to go back to the Enterprise LCMs or call Enterprise crisis lines.

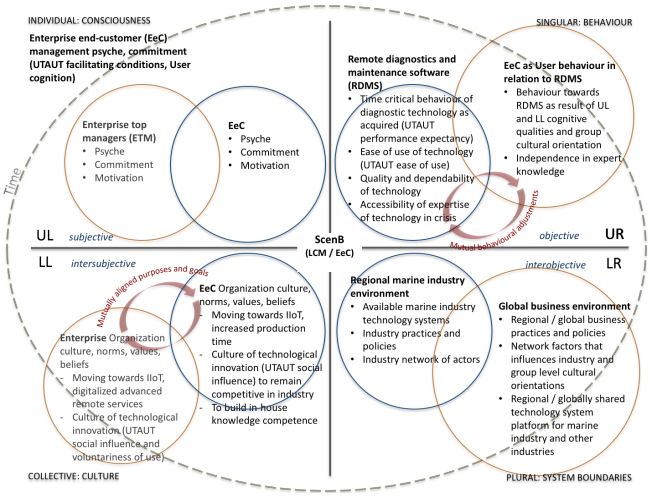


Figure 4. SFL pronoun system perspective of UTAUT Users for ScenB: Enterprise end-customer

The desire to be independent is due to EeC corporate sensitive and private data/information where it is not appreciated if Enterprise engineers can login remotely on unsecured internet connections to access EeC corporate systems data. EeCs are not

welcoming of what they consider to be an intrusion of privacy. How can EeC trust that it is an Enterprise engineer and not a competitor who has logged in was one such question raised to an Enterprise LCM, even if confidentiality agreements have been signed. EeCs also tend to want to keep operating costs down, so that having a remote diagnostics and maintenance contract is deemed useful when the system is predictive rather than reactive in feedback. This allows for action to be taken prior to failure. A movement towards reuse and recycling of parts also lowers costs for EeC. But even as EeCs might be aware of what more advanced remote services might offer, the incompatibility with older, existing systems takes time to disinvest or incorporate with newer systems and services.

Table 1. SFL UTAUT constructs for technology acceptance for combined Users: Enterprise, Enterprise engineers and Enterprise end-customers

SFL defined Users ->	ETM (ScenA)	EEHQ (ScenA)	EELBU (ScenA)	RSCS (ScenA)	LCM/EeC (ScenB)
SFL UTAUT constructs					
Performance expectancy (singular subjective, 'I')	-	c√	c√	√	-
Performance expectancy (plural intersubjective, 'We')	√	√	√	√	-
Performance expectancy (singular objective, 'It')	√	c√	c√	c√	c√
Performance expectancy (plural interobjective, 'Its')	f√	f√	-	-	f√
Effort expectancy (singular subjective, 'I')	-	√	√	√	√
Effort expectancy (plural intersubjective, 'We')	√	√	√	√	√
Effort expectancy (singular objective, 'It')	c√	c√	c√	c√	c√
Effort expectancy (plural interobjective, 'Its')	f√	f√	-	-	f√
Social influence (singular subjective, 'I')	-	-	-	-	-
Social influence (plural intersubjective, 'We')	√	√	√	√	√
Social influence (singular objective, 'It')	c√	c√	√	√	c√
Social influence (plural interobjective, 'Its')	f√	f√	√	√	f√
Behavioural intention (singular subjective, 'I')	√	√	√	√	√
Behavioural intention (plural intersubjective, 'We')	c√	c√	√	√	c√
Behavioural intention (singular objective, 'It')	c√	c√	c√	c√	c√
Behavioural intention (plural interobjective, 'Its')	f√	f√	√	√	f√
Use behaviour (singular subjective, 'I')	-	-	-	-	-
Use behaviour (plural intersubjective, 'We')	-	-	-	-	-
Use behaviour (singular objective, 'It')	c√	c√	c√	c√	c√
Use behaviour (plural interobjective, 'Its')	f√	f√	f√	f√	f√
Facilitating conditions (singular subjective, 'I')	c√	c√	c√	c√	c√
Facilitating conditions (plural intersubjective, 'We')	c√	c√	c√	c√	c√
Facilitating conditions (singular objective, 'It')	c√	c√	c√	c√	c√
Facilitating conditions (plural interobjective, 'Its')	f√	f√	f√	f√	f√

Users Key
ETM = Enterprise top management, EEHQ = Enterprise engineers, Headquarters
EELBU = Enterprise engineers, Local Business Unit,
RSCS = Remote services customer service personnel
LCM/EeC = Product Life Cycle Managers (reflecting Enterprise end-customers)

5.3. Combined SFL UTAUT Users Findings Summary

Table 1 summarises the findings of the SFL pronoun system defined Users, in relation to the six broad UTAUT category constructs for technology acceptance when placed into the SFL pronoun system quadrants.

Factors such as 'gender', 'age', 'experience' and 'voluntariness of use' are incorporated into the more specific pronoun system quadrants. The factors of 'gender', 'age' and 'experience' were accounted for in respondent profile through the primary requirement that all respondents were knowledge experts in their field regardless of gender or age. The Enterprise emphasises it is an equal opportunity employer. The combined Users include perspectives from Enterprise top managers, Enterprise engineers, remote services customer service personnel and Enterprise end-customers who feedback directly to Enterprise Product LCMs. Table 1 indicates which technology acceptance constructs are relevant (by '√'), which are critically relevant (by 'c√') and those that will be relevant in future (by 'f√'). UTAUT constructs with an indeterminate answer or that are non-relevant are indicated by '-'.

In overview, Table 1 indicates that what was defined as 'relevant', 'critically relevant' and 'future relevant' as UTAUT construct, depended upon the User perspective, as helped defined by the SFL pronoun system. In keeping with organization area of expertise, that the Enterprise defines itself as world leader in innovative technology, what was deemed as critically relevant for all Users was SFL UTAUT construct Effort expectancy (singular objective, 'It'), which meant that the remote services technology was expected to work for both the Enterprise and its end-customer. The technology was produced to be end-customer user friendly and accessible, so that learning input from the end-customer, whilst necessary, would be kept at minimal level. Critically relevant was UTAUT facilitating conditions for all SFL pronoun system perspectives as Users. This could be due to the fact that without adequate motivation from Individual level to conducive environment at Systems level, the technology would not be able to survive or evolve in the human technological ecosystem. Facilitating conditions also had to be right if a future (thus future relevant, 'f√') remote services standard platform for regional and global use is to be built / implemented. Its development will depend upon Enterprise commitment as well as customer demand and supporting industry policies and practices.

6. Study Limitations and Future Research

The strengths of this study are also its limits. This includes (i) using the Enterprise as case example unit of analysis and Enterprise only respondents, and (ii) using the SFL pronoun system in offering of a plurality of perspectives to UTAUT User constructs. Pertaining to strength and limitation (i), Enterprise as case example provides this study with a cohesive organization environment in which research method theory and analysis can be done. What has been assumed are shared corporate cultures and shared corporate strategies. The single organization as case example might in that sense crave comparative studies to be made, perhaps by definition of industry sector, even across industry sectors, comparing remote services technologies. Still, while single organization shared values has shown to be mostly true, it was not true that the Enterprise and its LBUs were homogeneous

in their workflows. This might have proved to be a challenge with regards to placing all Enterprise respondents into the majority ScenA Users. This is because it reflects that ScenA Users share the same User experience, when in effect, there are variations of context of use for ScenA system Users, which are not fully accounted for in this study due to that the focus of this study was to illustrate how SFL could be used to unfold different versions of Users usually statically defined when applying UTUAT models. The capturing of User nuances in remote services could in this case, be a subject area for future research, in view that remote services would be a field that grows in application with increased digitalisation of industry processes, beyond the marine sector into robotics and process automation. Since language is still its own metalanguage, SFL could be a means of a cohesive theory and framework towards a general systems theory perspective of studies on technology innovations and combined product-services. There were practical challenges to obtaining data for this study that included time coordination for respondents to be present during the field study and for shadowing purposes.

Pertaining to strength and limitation (ii), the use and application of the SFL pronoun system, while offering an unfolding of perspectives on any subject of study, is also reliant on clear definition of research purpose and design on the part of the researcher. Relativity of perspective *x,* will need to be defined, in relation to point of interest *y,* in context *[n+...]* in order to be effective in use. While the SFL pronoun system framework opened up various User definitions for UTAUT and made troubleshooting easier, in the sense that the SFL UTAUT defined User could be exactly pinpointed for unease of use of technology for example, what still remained were the working processes of communication between Enterprise business units and their end-customer on how to improve remote services efficiency. It could also be said that the SFL pronoun quadrant approach might not be too appreciated with a management or practitioner audience because the fluidity of perspectives as exacting as it might be, is also confusing, depending on audience. The remote services customer service personnel for example, who were the individuals meeting crisis calls from end-customers had little concern for how much commitment top management had towards building a standardised internal remote services platform for the future. They were rather more concerned that they could pass on the crisis call to an immediate available remote services engineer.

7. Conclusion

The purpose of this study is to bridge research methodology across different disciplines. In particular, it is to apply a language science model of analysis the SFL, unto a model of technology use and acceptance, UTAUT. This was done because a practical need arose to address multiple users of technologies in multiple work contexts in a single framework of analysis. The need for a plurality of perspectives of defined User (i.e. *which* user of technology?) arose in a case example of an Enterprise needing to develop a shared remote services platform for employees (mostly engineers) as well as end-customers. In such a context, both Enterprise engineers and Enterprise end-customer were Users of this technology system. UTAUT in its original application, is a model of technology use and acceptance that renders a linear statically defined User to a single context of use. Because language is inherently human and we use language to express and encode many internal and external experiences and world views, this study turned to SFL as functional language theory and framework of data analysis with the questions of (RQ1) *how* SFL can be applied to the UTAUT model so that the UTAUT model can be simultaneously applied for differently defined Users in different contexts and (RQ2) *what* contributing factors can be established by use of SFL as language theory and framework that broadens the applicability of UTAUT into various contexts? To that end, in answer to RQ1, this study has illustrated how the architecture of language contains in itself a myriad of language systems, of which the pronoun system is one. The SFL pronoun system unfolds primordial perspectives of Agency and Actorship set in both group and environmental context. In answer to RQ2, this unfolding of perspectives by applying the SFL pronoun system is what allows for the plurality of views of the defined Users of UTAUT constructs. The plurality of User views in different contexts when combined with UTAUT constructs allows for the immediate identification of disjunctive views occurring in the business workflows, when an Enterprise engineer believes the system to be user friendly but when the Enterprise end-customer does not. Or when two Enterprise engineers have different User experiences of the same system. This SFL UTAUT model allows for such gap in knowledge identification, and it gives those who work in the context a chance to reconcile these differences in opinion and work towards closing the knowledge gaps. The fluidity of the SFL pronoun system that also maps different types of knowledge zones also means that researchers can use the pronoun quadrant model to visualise research design perspective and address gaps in knowledge.

The application of the SFL pronoun system in complement to the UTAUT model is novel in research methodology. The resulting combined SFL UTAUT theoretical construct and the subject of study of remote services could be better developed by similar type multi-enterprise studies or comparative multi-enterprise type studies. And perhaps what the SFL UTAUT construct does not and cannot do, is to address the actual communication patterns between Actors and their surrounding context. Miscommunication is identifiable by the SFL UTAUT construct, but the act of improving on communication across different Enterprise business units and between their end-customers remains very much a human cognitive process.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

This study is funded by Stiftelsen Riksbankens Jubileumsfond of Sweden, as part of its Flexit (2015-18) Program, RMP15-0954:2. The purpose of the main project is to investigate linguistic methodological applications of analysis to user acceptance and use of technology.

References

[1] Fishbein, M., Ajzen, I., Belief, attitude, intention and behavior: An introduction to theory and research, Reading, MA: Addison-Wesley Publishing Company, 1975

[2] Ajzen, I., "The theory of planned behaviour", Organ Behav Hum Dec, 50(2), 179- 211, 1991. http://dx.doi.org/10.1016/0749-5978(91)90020-T

[3] Davis, F. D., Bagozzi, R. P., Warshaw, P. R. "User acceptance of computer technology: A comparison of two theoretical models", Manage Sci, 35, 982–1003, 1989. http://www.jstor.org.ezproxy.ub.gu.se/stable/2632151

[4] Venkatesh, V., Bala, H., "Technology acceptance model 3 and a research agenda on interventions", Decision Sci 39(2), 425-478, 2008. https://doi.org/10.1111/j.1540-5915.2008.00192.x

[5] Taylor, S., Todd, P.A. "Assessing IT usage: The role of prior experience", MIS Quart 19(2), 561-570, 1995. http://www.jstor.org/stable/249633

[6] Blackwell, C., Lauricella, A., Wartella, E., Robb, M., Schomburg, R., "Adoption and use of technology in early education: The interplay of extrinsic barriers and teacher attitudes: The interplay of extrinsic barriers and teacher attitudes", *Comput Educ,* 69, 310-319, 2013. https://doi.org/10.1016/j.compedu.2013.07.024

[7] Haas, S., Senjo, S., "Perceptions of effectiveness and the actual use of technology-based methods of instruction: A study of California criminal justice and crime-related faculty", J Crim Just Educ, 15(2), 263-285, 2004. https://doi-org.ezproxy.ub.gu.se/10.1080/10511250400085981

[8] Thakur, R., "Customer adoption of mobile payment services by professionals across two cities in India: An empirical study using modified technology acceptance model", Bus Perspect Res 1(2), 17-30, 2013. https://doi-org.ezproxy.ub.gu.se/10.1177/2278533720130203

[9] Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., "User acceptance of information technology: Toward a unified view", MIS Quart 27(3), 425-478, 2003. http://www.jstor.org.ezproxy.ub.gu.se/stable/30036540

[10] Barelka, A., Jeyaraj, A., Walinski, R., "Content acceptance model and new media technologies", J Comput Inform Syst 53(3), 56-64, 2013. https://doi-org.ezproxy.ub.gu.se/10.1080/08874417.2013.11645632

[11] Sykes, T.A., Venkatesh, V., Gosain, S., "Model of acceptance with peer support: A social network perspective to understand employees' system use", MIS Quart 33(2), 371-393, 2009. http://www.jstor.org.ezproxy.ub.gu.se/stable/20650296

[12] Maruping, L., Bala, H., Venkatesh, V., Brown, S., "Going beyond intention: Integrating behavioral expectation into the unified theory of acceptance and use of technology", J Assoc Inf Sci Technol 68(3), 623-637, 2017. http://dx.doi.org/10.1002/asi.23699

[13] Halliday, M.A.K., Matthiessen, C., Halliday's Introduction to functional grammar, 4th ed. Abingdon: Routledge, 2014

[14] Halliday, M.A.K., Meaning as choice. In, L. Fontaine, T. Bartlett and G. O'Grady (eds.), Systemic functional linguistics: Exploring choice, Cambridge University Press, p. 15-26, 2013.

[15] Capra, F., & Luisi, P. (2014). *The Systems View of Life: A Unifying Vision*.

[16] Capra, F. (2005). Complexity and Life. *Theory, Culture & Society, 22*(5), 33-44

[17] Martins, C., Oliveira, T., Popovič, A., "Understanding the Internet banking adoption: A unified theory of acceptance and use of technology and perceived risk application", Int J Inform Manage, 34(1), 1-13, 2014. https://doi.org/10.1016/j.ijinfomgt.2013.06.002

[18] Saadé, R. G., Kira, G., "Mediating the impact of technology usage on perceived ease of use by anxiety", Comput Educ, 49(4), 1189-1204, 2007. https://doi.org/10.1016/j.compedu.2006.01.009

[19] Shin, D., "Determinants of customer acceptance of multi-service network: An implication for IP-based technologies", Inform Manage 46(1), 16-22, 2009. https://doi.org/10.1016/j.im.2008.05.004

[20] Dalcher, I., Shine, J., "Extending the new Technology Acceptance Model to measure the end user information systems satisfaction in a mandatory environment: A bank's treasury", Technol Anal Strateg, 15(4), 441-455, 2003. https://doi-org.ezproxy.ub.gu.se/10.1080/095373203000136033

[21] Krishnaraju, V., Mathew, S., Sugumaran, K., "Web personalization for user acceptance of technology: An empirical investigation of E-government services", Inform Syst Front, 18(3), 579-595, 2016. https://doi-org.ezproxy.ub.gu.se/10.1007/s10796-015-9550-9

[22] Lee, J-H., Song, C-H., "Effects of trust and perceived risk on user acceptance of a new technology service", Soc Behav Personal, 41(4), 587-598, 2013. http://dx.doi.org/10.2224/sbp.2013.41.4.587

[23] Teo, T., "Modelling technology acceptance in education: A study of pre-service teachers", Comput Educ, 52(2), 302-312, 2009. https://doi.org/10.1016/j.compedu.2008.08.006

[24] Šumak, B., Heričko, M., Pušnik, M., "A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types", Comput Hum Behav, 27(6), 2067-2077, 2011. https://doi.org/10.1016/j.chb.2011.08.005

[25] Strudwick, G., McGillis Hall, L., "Nurse acceptance of electronic health record technology: A literature review", J Res Nurs, 20(7), 596-607, 2015. https://doi-org.ezproxy.ub.gu.se/10.1177/1744987115615658

[26] Pai, F-Y., Huang, K-I., "Applying the Technology Acceptance Model to the introduction of healthcare information systems", Technol Forecast Soc, 78(4), 650-660, 2011. https://doi.org/10.1016/j.techfore.2010.11.007

[27] Aggelidis, V. P., Chatzoglou, P. D.,"Using a modified technology acceptance model in hospitals", Int J Med Inform, 78(2), 115-126, 2009. https://doi.org/10.1016/j.ijmedinf.2008.06.006

[28] De Graaf, M., Ben Allouch, S., Van Dijk, J., "A phased framework for long-term user acceptance of interactive technology in domestic environments", New Media Soc, 1-22, 2017. https://doi-org.ezproxy.ub.gu.se/10.1177/1461444817727264

[29] Wagner, G. D., Flannery, D. D., "A quantitative study of factors affecting learner acceptance of a computer-based training support tool", J Eur Ind Training, 28(5), 383-399, 2004. https://doi-org.ezproxy.ub.gu.se/10.1108/03090590410533071

[30] Hasan, R. (2012). *Systemic Functional Linguistics: Exploring Choice* (Vol. 9781107036963). Cambridge University Press.

[31] Popova, O., Popov, B., Karandey, V. & Evseeva, M. (2015). Intelligence amplification via language of choice description as a mathematical object (Binary Tree of Question-answer System). *Procedia - Social and Behavioral Sciences, 214*, 897-905.

[32] Hasan, R., & Martin, J. (1989). *Language development: Learning language, learning culture* (Advances in discourse processes, 27). Norwood, N.J.: Ablex.

[33] Martin, J. R., "Meaning matters: A short history of systemic functional linguistics", WORD, 62:1, 35-58, 2016. https://doi-org.ezproxy.ub.gu.se/10.1080/00437956.2016.1141939

[34] Halliday, M. A. K. Language as social semiotic: The social interpretation of language and meaning, London: Edward Arnold, 1978.

[35] Whorf, B. L. Language, thought and reality: Selected writings, Cambridge, MA: MIT Press, 1956.

[36] Butler, C., "An ontological approach to the representational lexicon in Functional Discourse Grammar", Lang Sci, *34*(5), 619-634, 2012. https://doi.org/10.1016/j.langsci.2012.02.004

[37] Bartlett, T. "I'll manage the context": context, environment and the potential for institutional change. In, L. Fontaine, T. Bartlett and G. O'Grady (eds.), Systemic functional linguistics: Exploring choice, Cambridge University Press, p. 342-364, 2013.

[38] Bourdieu, P., Language and symbolic power, Cambridge: Polity Press, 1991.

[39] Wilber, K., A theory of everything, Boston: Shambhala, 2000.

[40] Butler, C., "Criteria of adequacy in functional linguistics", Folia Linguist, 43(1), 1-66, 2009. https://doi-org.ezproxy.ub.gu.se/10.1515/FLIN.2009.001

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 311-320 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Protecting Private Data using Improved Honey Encryption and Honeywords Generation Algorithm

Thanda Win*, Khin Su Myat Moe

Yangon Technological University, Department of Computer Engineering and Information Technology, 11010, Myanmar

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 02 October, 2018
Online: 14 October, 2018

Keywords:
Private data
Honey encryption
Honeywords
Distribution Transforming Encoder

ABSTRACT

Nowadays, many companies and organizations use various encryption techniques to secure their private data before send it through the unsecure network. However, many attackers try to store private data by using various attacks. Most of the organizations use password based encryption algorithm (PBE) to protect their private data. However, the existing PBE methods are vulnerable in brute-force attacks because of user-generated weak or repeated passwords. The problem of weak password based encryption algorithm prompted us to introduce honey encryption (HE). Honey encryption and honeywords help to minimize this vulnerability of password based encryption algorithm and they can prevent the text based messages from brute face attacks in order to make messaging communication more secure and efficient. However, the conventional HE has message space limitation in Distribution Transforming Encoder (DTE) process and storage overhead problem in honeywords generation process. Therefore, our proposed honey encryption algorithm uses discrete distribution function in DTE process to solve message space limitation instead of using cumulative distribution function. Storage overhead and typo safety problem can overcome by using our proposed honeywords generation method in honey encryption algorithm. Furthermore, we also propose hashing and salting algorithm for securing password and key expansion. In this paper, we describe the case studies as a calculation of DTE and hashing and salting algorithm. And then, we design and implement the improved honey encryption mechanisms and honeywords generation algorithm in message transmission process. Finally, we show the comparison results using mathematical model in DTE process and security analysis of our system.

1. Introduction

Most of organizations over the world want to exchange their private messages secretly over the unsecure communication channel. So, they use end to end encryption services with various data encryption algorithm for their private data. One type of encryption process called the end-to-end encryption (E2EE) can read the messages which own the secrete keys [1]. Most of the organizations use password based encryption (PBE) algorithm because user can select and remember their key easily. However, the existing PBE algorithms are weak because hackers can get the message easily if they find the keys using the various attacks. Among them, brute force attack is the vulnerable to the applications and websites using encryption algorithm because these attacks attack the password files in the database until the real key or passwords find [2]. So, many researchers try to prevent the brute force attacks by using various methods.

The two countermeasure methods to brute force attacks have been applied by many modern systems. The first method is to enhance the time complexity for attackers. When the time complexity increases, the attackers take many computation times for cracking the passwords and many systems protect the data by increasing key length pseudorandom number generator (PRNG). Therefore, many hashing algorithm such as MD5 and SHA256 are invented for improving computations times [3]. However, the brute force attack can reveal the key within the polynomial time. So, the honeywords generation methods are used in many cryptosystem for protecting the keys. The attackers can face the difficulties and can't stole the password file if the system uses honeyword generation method that stores the honeywords with the real passwords in the password files. The purpose of decoy passwords or honeywords are used for the attackers that attack the hashing password file using various attacks especially brute force attacks. So, the existing honeywords generation methods generate

*Thanda Win, Yangon , +959797102699, thanda80@gmail.com

at least 20 honeywords for one real password [4]. So, the existing method meets the storage overhead problem.

The second method is statistical coding scheme that can produce false plaintext messages [3] using ASCII code table. But, the attacker can easily know this false message is not the real message because these messages are meaningless messages. HE is an encryption algorithm that can produce a ciphertext which when decrypted by any wrong key by the brute force attacker, then it looks like meaningful fake plaintext which is not the original message. So this encryption provides security against brute force attack. Distribution transforming encoder (DTE) is the main part of HE and it performs the mapping of the message space into the seed space of binary bits string. According to the probabilities of a message in the message space, it maps the message to a seed range using DTE process in a seed space randomly. Then the resulting seed space is XORed with the key to get the ciphertext. For decryption, the ciphertext is XORed with the key and the seed is obtained. Then DTE uses the seed location to map it back to the original plaintext message. Even if the key is incorrect, the decryption process produces a honey message from the message space and thus can deceive the attackers [4]. In this existing HE algorithm, the message space can't accept more than four messages.

This paper described three contributions. Firstly, we report new DTE process for overcoming message space limitation problem in HE algorithm. And then, we design and implement the improved honey encryption system using new DTE process and apply the concept to message transmission application. These applications are based on uniformly distributed message spaces and the symmetric encryption mechanism. We also propose the new honeywords generation algorithm and hashing algorithm. Secondly, we apply two countermeasure attacks such as honey encryption and honeywords generation algorithms in message transmission process. Thirdly, we discuss comparison results of our proposed system and existing systems.

2. Related Works

For protecting the brute force attack, we firstly described two different methods such as hashing algorithm and statical code scheme using ASCII code table. However, these two methods can't fully against the brute force attacks. So the honeywords generation method and honey encryption algorithm are reported in the introduction. Although two methods can protect the brute force attack, they have weakness in their process. In this part, we describe the weakness of password hashing and ASCII code table. And then, we described the limitation problem of honeywords generation methods and honey encryption process as follows.

In the past year, D. Gross described the dangerous hashing passwords that are used in the various websites such as Evernote, Yahoo, LinkedIn and many other websites [5]. Due to the password cracking technique proposed by M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, the hashing password can easily know by the attackers [6]. P.G. Kelley, S. Komanduri, M.L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L.F. Cranor, and J. Lopez reported most of the passwords of users are weak and easy guessed password. So, their proposed method can easily know the hashing passwords of user [7]. The attackers can get three billion guesses per second if the system using some hashing algorithm [8]. Therefore, passwords are stored using hashing algorithm in the database are not against the various attacks.

S. Kharod, N. Sharma and A. Sharma proposed a hashing method using salting and differential masking that can provide the better security of the password with the faster time [9]. Although this method is better method for the passwords with better processing time, it can't strongly protect the brute force attack.

Above the description, hashing algorithm that uses for increasing computation time and producing false plaintext messages are not fully protect the various attacks especially brute force attack. Therefore, honeywords generation algorithm appeared in 2013 and honey encryption described in 2014.

In order to improve the safety of the hashed passwords, A. Juels and R. L. Rivest suggested a new method called honeywords generation method. They maintain the honeywords with the real password for each user's account by creating honeywords. The attacker cannot classify which password is real password if he gets inversion file of hashed password or honeywords. The auxiliary server or honey checker can classify the real password and honeyword for login process and will set off an alarm to categorize between the honeywords and real password [10]. However, this method can cause typing mistake of users during entering of password because honeywords generation method create the similar honeyword with the real password. The main weakness of this method is storage overhead problem.

Due to the development of graphical processing unit (GPU), the adversary can solve the hashing password files. So, Z. A. Genc, S. Kardas and M. S. Kiraz proposed the new method by keeping the decoy password and real password into the database. In this work, the author described the honeychecker for testing entering the false passwords to the system. Although getting the password file and coverting the hash code into the password, he can't enter the system without passing the honeychecker [11]. Therefore, this method can against the attacks of attackers, it meets the storage overhead problem.

For making less the existing problem of honeywords generation method, N. Chakraborty and S. Mondal was created the honey circular list algorithm. In this paper, the authors described circular list for storage of hashing password. When the attacker gets the password file, he cannot compute the distance of this password and he cannot login into the system. The honey circular list method can make less the storage problem of earlier existing algorithms [12]. However, this method can't almost solve the problem of honeywords producing methods.

S. R. Shinge and R. Patil presented a data encryption and decryption that uses values in ASCII code table. In this ASCII code table encryption approach, the plaintext messages and encryption keys are converting into the string using ASCII values [13] for producing ciphertexts. If an adversary stoles the key file and then he tries to decrypt the ciphertext use one key in this key files, he gets the meaningless ciphertext if he uses the false key. So, the attacker can know easily he uses the false key in this process.

A new encryption technique that can protect brute force attack called honey encryption (HE) algorithm was proposed by N. Tyagi, J. Wang, K. Wen and D. Zuo. When the attacker tries to get decrypted data by using brute force attack, that technique

can deceive the attacker and can give bogus meaningful messages. A new encoding and decoding scheme called a distribution transforming encoder (DTE) are used by HE and it has the limitation for assigning the plaintext messages [4]. The DTE in that system has limitation for placing plaintext message into the seed space.

According to the previous works, the existing system meets the storage overhead and message space limitation problems. Therefore, we propose a system to overcome message space limitation and storage overhead problem. This honeywords generation system can make less storage cost compared to the existing honeywords generation algorithm. Moreover, it can easily overcome the typo safety problem. For securing password files, we use new hashing and salting algorithm. The hashing and salting time is faster than the existing MD5 hashing algorithm and hashing algorithm using salting and differential masking . Finally our system using new honeywords generation and hashing algorithm, new DTE process can overcome message space limitation problem using DTE compared with the existing honey encryption algorithm.

3. Proposed Methodology

This part reports about the methodology of the proposed techniques such as the user authentication and login process, honeywords generation method, hashing and salting algorithm, distribution transforming encoder and honey encryption algorithm.

3.1. User Authentication and Login Process

In any login procedure, when the user is a new member, he needs to make the registration process. Otherwise, the user enters the system with his username and password. After making the registration process, the system produces the honeywords for this member using honeywords generation algorithm and the sweetwords that includes honeywords and real password are kept into the password file using our proposed hashing and salting algorithm. The registration process and login process are shown in the Figure 1(a) and 1(b). After the registration process, he can login into the system with his password.

The server detects the login password whether it has in the database or not. If the password exists in the database, the password is sent to the honeychecker by the server for categorization of honeyword and actual password. When the login password doesn't equal to the user's real password that stores in the database, administrator can get the alarm message from the honeychecker. Otherwise, the user can enter into the system successfully.

3.2. HoneyChecker

Honeychecker is an auxiliary secure server that can be used for categorisation of sweetwords including real passwords and false passwords. The honeychecker stores secret information such as the index of real passwords and the hashing passwords. The main server can communicate with the honeychecker when a user enters the system. When the false password or honeyword is entered into the system, the system administration can know immediately by raising an alarm signal from the honeychecker. In our system, the honeychecker performs two main processes. The first process is to classify the honeywords or real password when the user logging the system. The second is to send an alarm message to the administrator if the attacker enters with the honeywords.

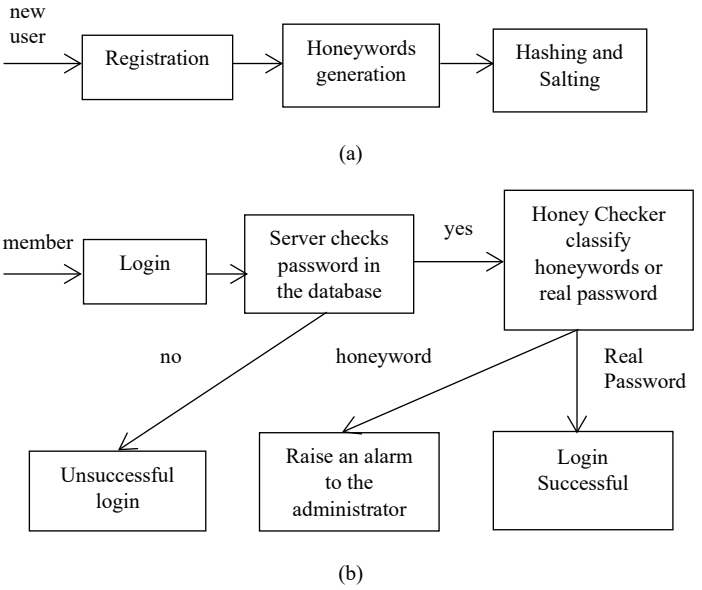


Figure 1: (a): User Registration Process (b): User Login process

3.3. Honeywords Generation Method

For every web application, passwords are notoriously weak authentication in the authentication process because users choose easily guess passwords by the attackers that apply dictionary, rainbow table, brute force attacks etc.. Therefore, as many technologies such as graphic processing unit (GPU) are developing, the attackers can get the user's password and can crack easily hashing password without consuming many times. So, the real passwords and the honeywords that are generated from our proposed honeywords generation algorithm are kept using our proposed hashing and salting algorithm for each account in the database by the system. The purpose of honeywords generation methods generates the honeywords to deceive the attackers [14]. However, the existing honeywords generation algorithms have the storage overhead problem.

Table 1. Example of password file P1 in main server

user name	honeyword index set
David	(21,35,40,55)
Suzy	(18,35,40,55)
Jane	(18,21,40,55)
Joy	(18,21,35,55)
August	(18,21,35,40)

Table 2. Example of password file P2 in honey checker

real password index	hashing password
18	763f7b09c8ac104bd
21	3e8fdfcd7f3615428b0c
35	7e8fcb6b2bc
40	365b4f15a8d445d
55	984d078ac47c2c82987d19

So, we proposed the new honeywords generation algorithm for making less the storage cost. In our proposed algorithm, we denote the member's passwords of our system as honeyword without creating and storing the individual honeywords in the database. Therefore, we create two password files P1 and P2 in the server's database and honeychecker's database. The password file P1 that includes username and honeyword index set in the main server's database. Real password index and hashing passwords including in password file P2 are kept by the system in the honeychecker's database. The table is sorted randomly according to the user registration position. The example storing of honeywords and real passwords are shown in table 1 and table 2.

The detail procedure of honeywords generation is described as follows.

Inputs:

1. username u_i and password p_i

2. After creating accounts, index list of real password and honeywords need to assign

Procedure

Step 1: User account registration and honeywords creation

a. P1 and P2 password files creation

P1 stores username and honeyindex set h_i, P2 keeps the index of real password r_i and the hash of the corresponding user's passwords $H(p_i)$

b. The list index set of honeywords h_i and real password p_i are created like the following

P1= {<u_i,h_i>, <u_2,h_2>, ----------,<u_i,h_i>}

P2= {<r_1,$H(p_1)$>, <r_2,$H(p_2)$>,-------, <r_i,$H(p_i)$>}

Step 2: User passwords and fake password storage:

a. P1 including username and honeyindex are set in main server
b. P2 including index of real password and hashing passwords are set in the honey checker

Step 3: HoneyChecker Classification

Set r_i, $H(p_i)$

Sets correct password index r_i for the user u_i

Set: Login password k

Check: r_i, k

If r_i is equal to the login password k, login process successful.

Otherwise, the honeychecker sends an alarm message to the administrator.

3.4. Hashing Algorithm

In our proposed system, honeywords and real passwords are stored into the database using our proposed hashing algorithm for providing better security of key or password with faster time. The detail steps of our proposed hashing algorithm are described the following.

Step 1: Change the user's entering password string into binary bits.

Step 2: Binary bits string are appended with padding bits 00000000.

Step 3: Change the string of binary bit 1 into 0 and 0 into 1.

Step 4: The changing string are make the XOR operation with the binary string by concatenation of 0's and 1's.

Step 5: The string rotates from right by 4 characters.

Step 6: The resulting binary string converts to hexadecimal string and stores into the password file that locates in the database.

The example of our hashing algorithm is shown as following. In this example user chooses her password as December30 and our proposed hashing algorithm converts into the hexadecimal string using the following steps.

Step1: Convert the input expression (password: December30) into binary string.

01000100 01100101 01100011 01100101 01101101 01100010 01100101 01110010 00110011 00110000

Step 2: Before making the password hashing, binary bits string are appended with padding bits 00000000 (green color).

00000000 01000100 01100101 01100011 01100101 01101101 01100010 01100101 01110010 00110011 00110000

Step 3: Change the string of binary bit 1 into 0 and 0 into 1.

11111111 10111011 10011010 10011100 10011010 10010010 10011101 10011010 10001101 11001100 11001111

Step 4: The changing string are performed the XOR operation with the binary string by concatenation of 0's and 1's.

String1:

11111111 10111011 10011010 10011100 10011010 10010010 10011101 10011010 10001101 11001100 11001111

String 2:

00001111 00001111 00001111 00001111 00001111 00001111 00001111 00001111 00001111 00001111

After XORing operation, the resulting binary string are described as follows.

11110000 10110100 10010101 10010011 10010101 10011101 10010010 10010101 10000010 11000011 11000000

Step 5: The resulting string rotates from right by 4 characters. So, first 4 characters are rotated to the right.

00001111 00001011 01001001 01011001 00111001 01011001 11011001 00101001 01011000 00101100 00111100

Step 6: The resulting rotate string converts to hexadecimal string and stores into the password file that locates in the database. So, the resulting hexadecimal string is as

follows.

Hexadecimal String - 7B49593959D929582C3C

3.5. Distribution Transforming Encoder

Distribution transforming encoder (DTE) process is the main key of honey encryption algorithm that assigns the message space M to the binary bits string of seed space S. Depending on the number of messages, the range of seed space assigned for distribution. The existing DTE process uses cumulative distribution function to map the message space into the seed space and it meets the message space limitation problem [4]. The DTE process includes encoding and decoding. In the encoding process, the DTE apply the discrete distribution function for assigning the plaintext messages M to the seed space S. The DTE process encodes the plaintext messages M to the range of seed space S is assigned randomly. On the other hand, the decoding process is harder than encoding process. In the decoding process, the seed space S decodes into the message space M using DTE process. One of our main contributions is to construct the DTE process using discrete distribution function to overcome message space limitation problem.

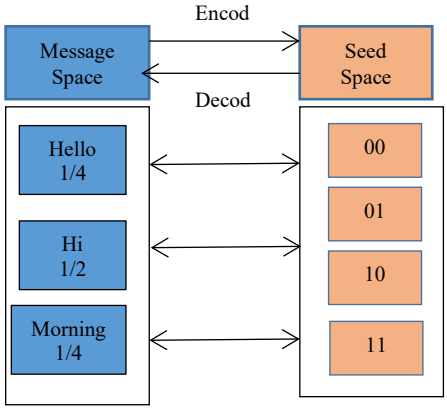


Figure 2: Basic Example of DTE Process

Figure 2 illustrates a basic example of DTE process. In this figure, message space consists of different elements in this case M={ hi, hello, morning}. Depending on the number of messages, the probability of each element using discrete distribution function is generated.

3.6. Honey Encryption Algorithm

Password-based encryption (PBE) and hashing techniques cannot fully against the brute force attacks since most of the users that use password as encryption key select the weak and repeatedly user-generated passwords. Honey encryption (HE) algorithm is one of the deceive algorithm to attackers who use brute force attack. The purpose of HE algorithm is to lure the attackers. HE generates the honey message through the brute force attack and deceives the attackers in a way that they cannot classify which messages are true messages. The core creation of existing HE is the distribution transforming encoder (DTE) that uses cumulative distribution function. Constructing the better DTE process can provide to get better HE algorithm [4].

The contribution of this paper consists of three main categories: honeywords generation, improved hashing algorithm and distribution transforming encoder. In our honeywords generation algorithm, storage overhead problem can be reduced. Our improved hashing algorithm is faster than existing MD5 hashing algorithm. Our improved DTE process that uses discrete distribution function can improve the message space limitation. Figure 3(a) and 3(b) show the basic encryption and decryption process of honey encryption algorithm.

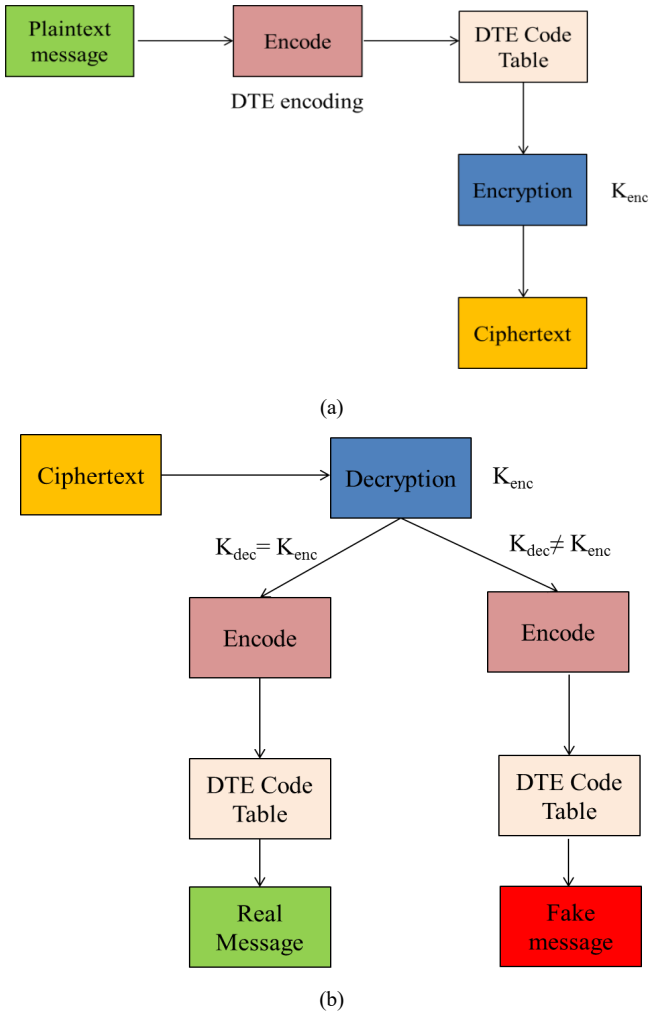


Figure 3 (a) Basic Encryption Process of HE Algorithm (b) Basic Decryption Process of HE Algorithm

The encryption process includes the plaintext message M encoding process using encoding code table and encrypting with encryption keys K_{Enc} to produce the ciphertext. The code table is made by using DTE process that use discrete distribution function. The sender sends resulting ciphertext to the receiver via the internet.

In the decryption process, the receiver decrypts the ciphertext with K_{Dec}. The decryption result decodes with the code table that equals the code table using in encryption process. If encryption key K_{Enc} and decryption key K_{Dec} are the same, the receiver can gets a *real* message that the sender sends. Otherwise, the receiver will get a bogus message that look like real message.

4. Proposed System Implementation

This section discussed about the flowchart of our proposed system, system design, example of the proposed system and mathematical model of the proposed system.

4.1. Flowchart of Proposed System

Figure 4 describes the detail flowchart of the proposed system.

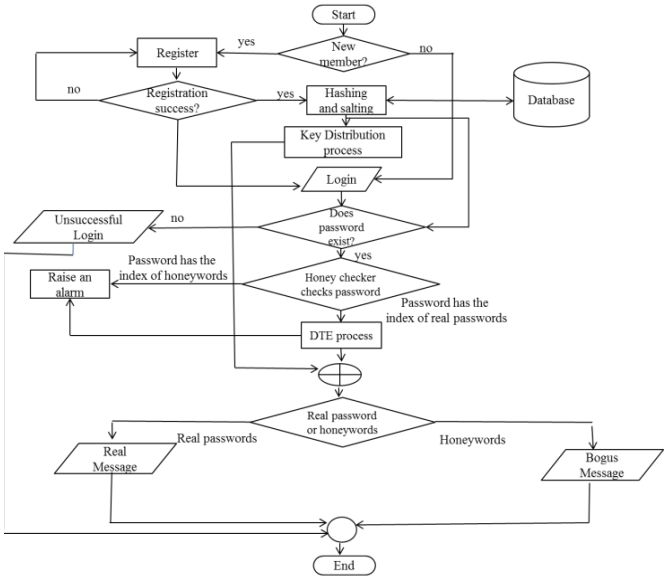


Figure 4: Flowchart of Proposed System

If a new user wants to enter this system, he needs to make the registration process at least one time. He doesn't need to make the registration process to enter the system if he is a member of the system. After making the registration process, he can login into the system with his username and password. If he enters into the system with his username and password, the system checks his password exists or not in the database. In this stage, he can get three conditions. Firstly, he can get unsuccessful login message from the system if his password doesn't have in the database. At that time, he needs to check his username and password if he is a member of this system. Otherwise, he needs to make registration process. His password will be sent by the server sends his password to honey checker for categorization of honeyword or sugarword if his password exists in the database. Secondly, if his password is real password, he will perform the DTE process and gets the real message from the sender's send. Otherwise, he will perform the DTE process and he will get the bogus message.

4.2. System Design of the Proposed System

The process of our proposed system includes DTE process and key or password distribution process. DTE process makes encoding and decoding of original plaintext message. DTE that uses discrete distribution function for encoding message space M also called plaintext message such as \$5000, \$10000, \$750, \$900, \$200 into the seed space S. The range of seed space denotes as binary bits. The number of plaintext message space depends on the directly proportional to the range of seed space. For placing the number of messages m in plaintext message M to seed space S_m, DTE requires discrete distribution function in order to overcome message limitation.

In keys or passwords distribution, honeywords generation and key expansion of this password file of honeywords and real passwords are considered. In the honeywords generation process, we generate the honeywords by denoting the other user's real passwords and stores as honey indexes for overcoming weakness of existing honeywords generation methods such as storage overhead problem and typo safety problem. By using the other's sweetword as decoy password, the adversary becomes complicated which one is correct password or honeywords. In this case, the password and honeywords are stored in the password file using our proposed hashing algorithm to provide better security. And then, the resulting password and honeywords are randomly placed into the seed space S_k using DTE process. After placing password file into the seed space S_k, the key expansion result S_k is XORed with the message distribution result S_m of DTE. Finally, the ciphertext message is successfully created. During the decryption process, the ciphertext is XORed with the secret key to get the seed space S_m. And then we decode this seed space S_m to get original plaintext message.

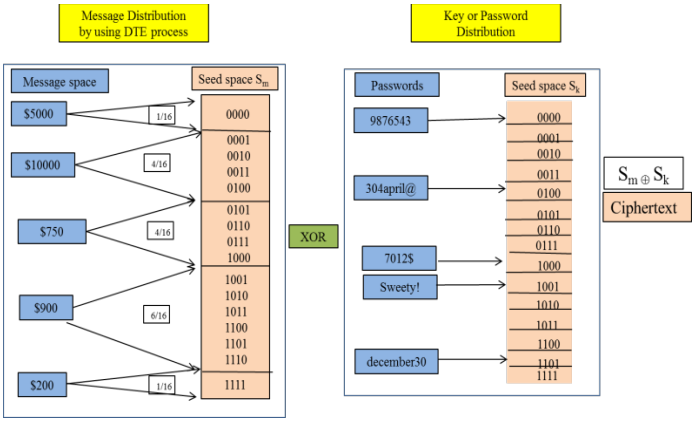


Figure 5: Proposed System Design

4.3. Example of Our Proposed System

Figure 5 describes the detail example design process of our proposed system. In this example, different messages such as M = {\$5000, \$10000, \$750, \$900, \$200} are denoted as message space. In this case, message space M has five elements. According to the number of messages, probabilities are assigned to each element. By calculating, the seed space range of message M is $2^{5-1} = 2^4 = 16$ and seed space is 4-bits binary strings. Depending on the results of probabilities values, different messages elements are mapped into the range of seed space by the system. For assigning the message space, the messages are ordered randomly. With this randomly ordering, the messages get different probabilities and the range of seed space will be different. The secrete key or passwords are K = {9876543, 30april@, 7012\$, Sweety!, december30}.

Using discrete distribution function, probability of the messages is {1/16, 4/16, 6/16, 4/16, 1/16}. So, the seed space S_m of the message "\$5000" is {0000} and "\$10000 is {0001, 0010, 0011, 0100}. And then, seed space S_m of \$750 is {0101, 0110, 0111, 1000, 1001, 1010} and \$900 is {1011, 1100, 1101, 1110}. Finally, the seed space S_m of \$200 is {1111}. According key distribution process, the seed space S_k of 9876543 is {0000} and

april30@ is {0100}. The next 7012$ is {1000} and Sweety! is {1001}. And then, the seed space S_k of december30 is {1101}.

In the encrypting process, Bob wants to send encrypting message "$5000" to Alice under the secrete key as april30@. So, we compute seed space S_m by using DTE process and then the result S_m of $5000 is (0000).Similarly, according to the key or password distribution, seed space S_k of password april30@ is (0100). And then, Bob sends the resulting ciphertext (0100) by XORing the seed space S_m and S_k to Alice.

The decryption process is harder than encryption process because it includes message recovery process. In this process, Alice tries to decrypt ciphertext C by using password april30@ of Bob. Firstly, Alice computes seed space of key distribution S_k and this S_k (0100) is XORed with the ciphertext (0100) to get seed space of message distribution S_m (0000). And then, he decodes S_m by using DTE process to get original plaintext message "$5000".

On the other hand, attacker John wants to get Bob message. So he attacks password file by using brute force attack. John gets password file but he doesn't classify which is Bob password. So, he randomly chooses password 9876543 and tries to decrypt Bob message. John computes seed space of 9876543 to get key distribution S_k (0000). And then he makes XOR process of ciphertext C and seed space S_k. The resulting seed space of message distribution S_m is (0100). And then, he decodes S_m by using DTE process to get Bob message. Finally, John gets "$10000" after decoding process of seed space of message distribution. In this process, he doesn't know which message is true message or decoy message. This encryption algorithm can make confusion for attackers and it can strongly protect brute force attack.

4.4. Design Methodology

The methodology proposed in this research work is based on Honey encryption algorithm using honeyword Generation Method. The above mentioned work is divided into various modules as follows:

- Registration: At the time of registration, entering password, the system converts password into hash values and store into database. Along with hash values the real password's hash is also store at specific random position. After entering encryption key for encryption and decryption, system will generate honeywords for deceiving the attackers.
- Honeyword Generation: Honeywords are generated from the real password and in case any hacker tries to hack into the account by guessing the password the main user is sent alerts in form of a mail or some message so he knows that somebody is trying to login using his or her account.
- Login: If a user wants to login into the system, the user will type username and password. If the password matched with the hash password, the user is authenticated.
- Message Encryption and Decryption: User can send encryption message via the communication channel using encryption and decryption key.
- Generating honey message or fake message: If the encryption key is not match to the decryption key, the system will generate the honey message or fake message for counterattacking the attacker.
- Admin Management: Admin has the privileges to control over the mechanism. Admin has the authorization to maintain the user accounts and messages management.
- Hacker: If the hacker tries to break the system and enters any honeyword then the alert is given to the Actual user. When the attacker gets the encryption key file and he tries to decrypt the ciphertext using the honeywords, he will get the honey messages.

4.5. Mathematical Model of Our Proposed System

The core process of HE uses DTE for assigning the message space M into the seed space S_m using discrete distribution function in our proposed system. The discrete distribution function is a branch of probability function. The values z in the sample space Ω of discrete probability f(z) theory must have between 0 and 1 and the total values of f(z) for all values z in the sample space Ω must equal to one. This discrete probability theory satisfies the following two properties [15].

1. $f(z)\varepsilon [0,1]$ for all $x \varepsilon \Omega$
2. $\sum f(z) = 1$

- DTE using Discrete Distribution Function

We denote S be seed space or sample space and the value of seed space is [0,1]. According to the discrete distribution function, we calculate to get one for the total value of probability P(z) for all message in seed space S. Let n be total number of messages, x be number of '0" observed, M be the message space that contains number of messages and m be message. The following examples prove that DTE using discrete distribution function satisfies the properties of discrete probability theory. In this example we use number of five messages M={m1,m2,m3,m4,m5}. Firstly, we calculate seed space S.

Number of messages n= 5

For Distribution transforming encoder,

Seed space $S_m = 2^{n-1} = 2^{5-1} = 2^{5-1} = 2^4 = 16$

S={0000,0001,0010,0011,0100,0101,0110,0111, 1000,1001, 1010, 1011, 1100, 1101,1110,1111}

By using Discrete Distribution Function,

m_1=1/16
m_2=4/16
m_3= 6/16
m_4= 4/16
m_5= 1/16

M	m1	m2	m3	m4	m5	Total
P(z)	1/16	4/16	6/16	4/16	1/16	1

- DTE using Cumulative Distribution Function

According to the cumulative distribution function, we should get one for the total value of probability P(z) over all messages in seed space S to satisfy the properties of probability theory [10]. The following example describes that DTE using cumultative distribution function cannot satisfy the probability theory because the total values of probability P(z) of message in seed space S doesn't get one [15]. In this example we use five messages M= {m1,m2,m3,m4,m5} and calculate the probability P(z).

Number of message n= 5

For Distribution transforming encoder,

Seed space $S_m = 2^{n-1} = 2^{5-1} = 2^4 = 16$

By using cumultative distribution function

m1= 1/16 [(1,1)]

m2= 2/16 [(1,2),(2,1)]

m3= 2/16 [(1,3),(3,1)]

m4= 3/16 [(1,4),(4,1),(2,2)]

m5= 2/16 [(1,5),(5,1)]

M	m1	m2	m3	m4	m5	Total
P(z)	1/16	2/16	2/16	3/16	2/16	10/16

5. Results and Discussion

Many security areas such as web page and application using internet service used PBE for protection of their data from various attacks. However, PBE can't solve problem of the various security attacks especially brute force attack because most of the passwords are not much stronger. Our proposed honey encryption algorithm and honeywords generation methods can against brute force attack and many other attacks by deceiving the attacks. Moreover, our proposed method overcome the weakness of existing honey encryption and honeywords generations algorithm. The detail analysis and comparison results are shown in the following.

5.1 Implementation and Experimental Results

In this section we implement our proposed methodology in web application by developing a simulated messaging application and the language used is python programming language.

To evaluate the performance of our proposed honey encryption algorithm, we need to consider the false messages. The testing of the performance was conducted to distinguish between the correct key and incorrect keys. In this testing, we encode the message using the correct key. Then, we decrypt the message using the various incorrect keys. For all the testing, our proposed system generates the fake messages that are different from the original messages. The Figure 6 shows the original plaintext that was transmitted.

An attacker trying several keys is presented with the version of Figure 7 and Figure 8. There is no way the adversary can tell the original message from the fake message even if he acquires the original message.

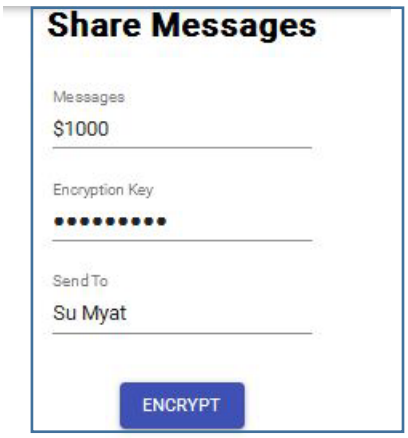


Figure 6: Original Plaintext Message

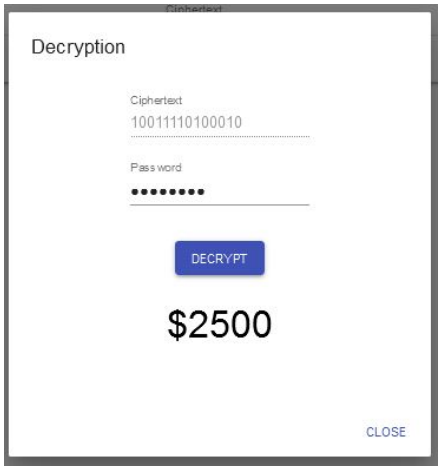


Figure 7: Intercepted Message acquired by Attacker using incorrect key

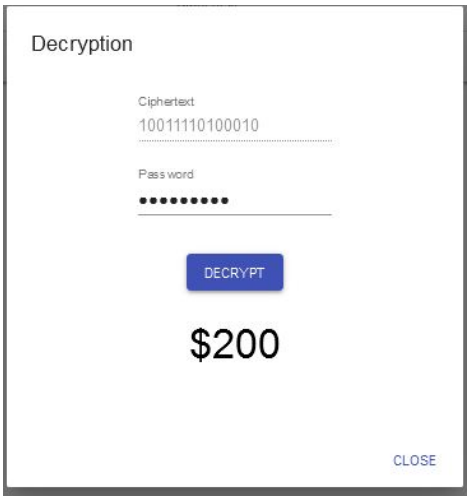


Figure 8: Intercepted Message acquired by Attacker using incorrect key

5.2 Comparative Analysis of Our Proposed System

Honey encryption (HE) algorithm is an interesting challenges algorithm and it plays an important role in various areas such as

business groups, banking system and computer users. Moreover, HE can protect from large number of attacks. Here it examines the comparative analysis of our proposed system and existing systems are described as follows.

1) Brute Force Attack

The web application is attacked by the attackers using the various attacks. Among them, the most dangerous attack for websites is brute force attack. The purpose of this attack is to steal the passwords and encryption key for obtaining the encrypted data. The attacker attacks the password file using the brute force attack and it finds the possible keys and password until the correct one is found. However, the attackers can't get the encryption key because of our proposed method. Although the password file in the database has been attacked by brute force attack, the attacker cannot classify which is real key or password if he gets password file. Moreover, attackers can take many times comparing into the existing security system because our system uses improved hashing and salting algorithm.

2) *DoS Attack*

The machines or network resources can be stopped and cann't be given any services to the users by the Denial of service (DoS) attack. Our proposed system is intended to reduce the DoS vulnerabilities. When a person tries to login into the system, the system automatically checks login of the user. If the attacker tries to login with his attacking password file, the system raises alarm for user and attack detected messages. And then, the system can block that particular attacker.

3) *Dictionary Attack*

Dictionary attack is one of the types of password guessing attacks and an attacker tries to login by cycling through combination of common words. Generally, this attack succeeds because most of the users choose weak or repeated passwords. However, our system can protect the passwords from the dictionary attack. When the attacker tries to login with randomly chooses an account with guessed password from password file, it can be two probabilities which are successful or fail. First is taking the password from the password file is correct and the next is getting the honeywords. If the attacker login with honeywords, the system raises an alarm for the user account and can detect this attack. Otherwise, if he enters with the real password easily, he can get the real message easily. But he cannot classify this message is real or fake messages.

4) *DTE Process*

The existing honey encryption algorithm and our proposed system denotes DTE as the main process. In this HE process, the message space that includes all plaintext messages is mapped into the seed space using DTE process. The existing DTE process uses cumulative distribution function for mapping the message into the seed space and it meets the message limitation problem. In our proposed system, DTE uses discrete distribution function and it can overcome the message limitation problem.

Table 3. Comparison Results of Existing Methods and Our Proposed Method

Techniques	Brute Force Attack	DoS Attack	Dictionary Attack	DTE Process
Password - based Encryption	Weak	Weak	Weak	-
Honey Encryption	Strong	Strong	Strong	Message Space Limitation
Our Proposed System	Strong	Strong	Strong	Overcome Message Space Limitation Problem

5.3 Comparison Results of Existing Honeywords Generation Methods and Our Proposed Honeywords Generation Method

In this section, we describe the comparison of the honeywords generation methods including our models in terms of flatness, typo safety and storage overhead.

1) *Flatness*

Our proposed system kept the list of s sweetwords that includes the real passwords and honeywords in the passwords files. There are two types of flatness such as perfectly-flat and approximately-flat for computing the security levels of honeywords generation methods. Depending on the types of flatness, we can decide which honeywords generation method is the better method for securing of our applications. Flatness computes that the attackers get how many times the real passwords. Let consider the attacker attacking the probability is w and this attacking probability compares the results of real password probability 1/s. When w>=1/s , the honeywords generation is approximately-flat and an adversary can guess the real password at random position. Otherwise, it said to be perfectly-flat if w<s [5].

Our proposed system is perfectly-flat and the attacker can't guess the real password position because we choose the honeyword from the other user's passwords. Our system has more advantages in security when the system includes more members.

2) *Typo Safety*

This is the important problem of existing honeywords generation algorithm. By creating the similar honeywords using honeywords generation algorithm, most of the users become typing mistake problem [5]. However, our proposed system can mostly reduce this typo safety problem comparing with the existing methods. Due to the using of other user's passwords as honeywords, the user can't miss with other user's passwords.

3) *Storage Overhead*

In the existing honeywords generation process, the algorithm generates at least twenty honeywords for deceiving the attackers. So, the attackers conflict and can't easily distinguish which one is real password or false password. But,

the system stores at least 21 passwords for one user in the database and the database becomes the storage overhead problem. In our proposed system, we denote s is sugarword that includes real password and honeywords. So, the honeywords become (s-1) because the user has one real password. Among the existing system, paired distance protocol (PDP) method is the best reducing storage overhead problem and it stores only random string of honeywords without storing the whole honeywords [7]. However, our proposed method can mostly reduce the storage overhead than the PDP algorithm because our system uses the indexes of member's password as honeyindexes.

Table 4. Comparison Results of Existing Honeywords Generation Methods and Our Proposed Method

No	Methods	Flatness	Typo Safety	Storage Overhead
1	CTD	1/k	Low	(s-1)*n
2	Password Model	1/k	High	(s-1)*n
3	Take a Tail	1/k	High	(s-1)*n
4	PDP	1/k	High	(1+RS)*n
5	Our Method	1/k	High	(1*n)

CTD- Chaffing by Tweaking

PDP- Paired Distance Protocol

s- Number of sweetwords in password file

n- Number of users

RS- Random String

5.4 Time Comparison Results of MD5 and Our Proposed Hashing Algorithm

Our proposed honey encryption process are designed by python programming language. This python programming language is flexible for all real world system and easy for the programmers to study. In this paper, we proposed three main contributions: improved DTE process using discrete distribution function, new honeywords generation method and the last is improved hashing algorithm for providing the better security of our proposed system with less time complexity comparing to the existing MD5 hashing algorithm. From the experimental results, the MD5 hashing algorithm takes more time than our proposed algorithm for converting the passwords into the hash code because our proposed system use simple binary operation.

Table5. Time Comparison Results of Existing MD5 Hashing Algorithm and Our Proposed Algorithm

		Time Complexity (Milliseconds)			
Length of Passwords (in character)		6	8	10	12
Methods	MD5	1740	2854	3260	3681
	Our Proposed Algorithm	31	46	62	78

In the following table, we uses the user's passwords such as "123456, hello123, student321, December3040" for the comparing the time results of other proposed hashing algorithm and existing MD5 hashing algorithm.

6. Conclusions

Most of the organizations try to against the brute force attack for protecting their private data. Honeywords generation algorithm can make the confusion for the attackers by storing the real password with the false passwords or honeywords in the password files. However, honeywords generation algorithm can meet the storage overhead problem if many of the users use this application. Honey encryption algorithm can protect the brute force attack by producing honey messages if the attackers use the honeywords. The existing honey encryption algorithm has the message space limitation problem and it can't send more than four messages by using cumulative distribution function. Our proposed honeywords generation algorithm and honey encryption algorithm can solve the storage overhead problem and message space limitation problem. Moreover, we proposed new hashing algorithm for securing our system and for getting the faster processing time compared to the existing MD5 hashing system.

References

[1] J.H. Saltzer, D. P. Reed and D.D. Clark, "Review for Paper: End to End Arguments in System Design" Polytechnic University, USA, 2006.

[2] K. S. M. Moe, T. Win, "Enhanced Honey Encryption Algorithm for Increasing Message Space against Brute Force Attack" in 15th Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) Conference, Thailand, Chaing Rai, 2018.

[3] A. Juels and T. Ristenpart, "Honey encryption: security beyond the brute-force bound," in Advances in Cryptology EUROCRYPT 2014, Springer, Berlin, Germany, 2014, vol. 8441 of Lecture Notes in Computer Science, pp. 293–310.

[4] N.Tyagi, J. Wang, K. Wen and D. Zuo, Honey Encryption Application, Computer and network Security, Springer, 2015.

[5] D. Gross, "50 million compromised in Evernote hack" CNN, 4 March 2013.

[6] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," In IEEE Symposium on Security and Privacy (SP), pp-162-175, 2009.

[7] P.G. Kelley, S. Komanduri, M.L. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L.F. Cranor, and J. Lopez, "Guess again (and again and again): Measuring password strength by simulating password- cracking algorithms," In IEEE Symposium on Security and Privacy (SP), pp- 523-537, 2012.

[8] M. Bakker and R. V.D .Jagt, "GPU-based password cracking. Technical report," Univ. of Amsterdam, 2010.

[9] S. Kharod, N. Sharma and A. Sharma, "An Improved Hashing Based Password Security Scheme Using Salting and Differential Masking, in 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Direction)," India, 2018.

[10] A. Juels and R. L. Rivest, "Honeywords Making Password Cracking Detectable" in ACM SIGSAC Conference on Computer & Communications Security, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 145–160.

[11] Z. A. Genc, S.Kardas and M.S. Kiraz, "Examination of a New Defense Mechanism: Honeywords," in 11th WISTP International Conference on Information Security Theory and Practice, Heraklion, Crete, Greece, 2017.

[12] N. Chakraborty and S. Mondal, "A New Optimized Honeyword Generation Approach for Enhancing Security and Usability," arXiv preprint arXiv: 1509.06094v1, 2015.

[13] S. R. Shinge and R. Patil, "An Encryption Algorithm based on ASCII Value of Data" International Journal of Computer Science and Information Technologies, vol. 5(6), 2014, pp-7232-7234.

[14] A. Juels and R. L. Revist, "Honeywords: Making Password Cracking Detectable" In MIT CSAIL, 2013.

[15] B. Addanki, Computer Physicist, "What is the difference between a probability density function and cumulative distribution function," 2015. http:// www.quora.com.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 318-326 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

A Comparative Analysis of two Controllers for Trajectory Tracking Control: Application to a Biological Process

Abyad Mohamed*, Karama Asma, Khallouq Abdelmounaim

LAEPT-URAC 28, Faculty of Science Semlalia, Cadi Ayyad University, Marrakech, Morocco

ARTICLEINFO

Article history:
Received: 01 June, 2018
Accepted: 04 August, 2018
Online: 20 August, 2018

Keywords:
Bioprocess
Constraints
Fuzzy observer
LMIs
Takagi-Sugeno Model
Trajectory tracking

ABSTRACT

The aim of the present work is to guarantee the trajectory tracking of a nonlinear biological process and compare two control approaches. The main objective of this work is to elaborate a fuzzy model and build a fuzzy controllers for a biological process by using the fuzzy Takagi-Sugeno. Two controllers are synthesized, the parallel distributed compensation control and optimal fuzzy linear quadratic integral control. In both cases, the physical constraints on the manipulated inputs are respected. In addition, the case with and without the observer is presented, where a fuzzy observer based control is used with unmeasurable premise variables. Finally, the performances and the effectiveness of both the modeling and the control are demonstrated via simulations.

1 Introduction

Nowadays, the biological processes become one of the important industrial processes thanks to their advantages, such as the treatment of organic substrates, protein production and the production of ethanol gas etc. However, their modeling and control form a real challenge problem for both control engineers and theorists, where this kind of systems are characterized by strong variations of system parameters and unknown kinetics owing to the time-varying characteristics and multiple interactions generated by the living microorganisms [1, 2]. Therefore, we obtain a highly nonlinear system. The motivation of this work is to linearize the model and benefit from linear theory control and to try to develop a nonlinear control, which is very difficult in this case. Also to use the Takagi-Sugeno (T-S) model, furthermore, the proposed controllers can be applied to the real process. It only needs to identify a T-S model from experimental data. T-S approach has been recognized as an effective tool for handling the previous difficulty.

There are different techniques for controlling the bioprocess using Takagi-Sugeno models, such as optimal fuzzy linear quadratic regulators for discrete-time [3], a fuzzy integral controller to force the switching of a bioprocess between two different metabolic states is treated in [4], an internal model control design strategy is developed for a particular Continuous Stirred-Tank Reactor (CSTR)[5]. A PID and fuzzy controller are proposed in [6] to stabilize the CSTR around the equilibrium point, where the authors consider only one input, which is not the case in practice. Also, the case of uncertain Takagi-Sugeno system is treated in [7], where an observer with unmeasurable premise variables and unknown input is considered for a wastewater treatment plant. In addition, the predictive control based on fuzzy observer is studied for a sludge depollution bioprocess in [8, 9], in this framework one can cite [10, 11, 12]. Furthermore, the modeling and the control of bioprocess based on neural network approach is treated in [13, 14]. In the same spirit, a nonlinear model autoregressive with exogenous input model predictive control is developed in [15] to control the fermentation process. Also, an integral backstepping control law is developed in [16] for controlling the dissolved oxygen level for bacteria fermentation.

The problem treated in this paper is how to model and control the biomass growth process, ensuring the trajectory tracking while taking into account the following constraints:
- The mathematical model is nonlinear and not affine in control.
- The variables control present the physical constraints,

*Corresponding Author: Abyad Mohamed, Faculty of Science Semlalia Marrakech , Morocco, abymedlmp@gmail.com

https://dx.doi.org/10.25046/aj030432

which make the computation of the control gains difficult.
- The full system states are not measurable.

The present paper has two goals, the first is to build a fuzzy model of biological process based on Takagi-Sugeno tool, especially the nonlinearity sector methods. The second is to ensure the tracking trajectory of the desired outputs using two approaches: the Parallel Distributed Compensation (PDC) [17] and the Linear Quadratic Integral (LQI) control. Where the strong physical constraints on the inputs [18] are taking into account. In addition, the proposed controllers are compared. The stability conditions are formed in the Linear Matrix Inequalities (LMIs) terms.

This paper is organized as follows: Section 2 presents the description of the Takagi-Sugeno modeling. Section 3 describes the parallel distributed compensation control. Then, one can address the fuzzy output tracking control problem in section 4 and we show that it can be solved by using two methods: the PDC technique and the optimal linear quadratic control. Section 5 describes the controller design based on fuzzy observer with unmeasurable premise variables. Section 6 introduces the proposed biological process. Finally, the simulation and the discussion of the obtained results are given to compare the proposed controllers.

2 Takagi-Sugeno Fuzzy Model

In order to extend the existing approaches of control and observation for linear to nonlinear system, Takagi and Sugeno have proposed a fuzzy dynamic model to represent this kind of system. The T-S fuzzy model is a set of linear models connecting via membership functions. To build the T-S fuzzy model, three methods exist in the literature [17]: The black box identification, the linearization method and the nonlinearity sector methods. The third method gives an exact T-S representation of nonlinear system without information loss.

Consider the following nonlinear system:

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + g(x(t))u(t) \\ y(t) &= Cx(t) \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ is the input vector, $y \in \mathbb{R}^q$ represents the output measurement vectors and $C \in \mathbb{R}^{q*n}$ is the output matrix. In addition, $f(.)$ and $g(.)$ represent the nonlinear functions.

The T-S fuzzy model uses a set of fuzzy if-then rules, which represent local linear input-output relations of a nonlinear system. The ith rule of the T-S model given as follows:

Rule i:

if $z_1(t)$ is $F_1^i(z_1(t))$ **and** $z_2(t)$ is $F_2^i(z_2(t))$ ... **and** $z_p(t)$ is $F_p^i(z_p(t))$

$$\textbf{Then} \begin{cases} \dot{x}(t) = A_i x(t) + B_i u(t) \\ y(t) = C_i x(t) \end{cases} \tag{2}$$

Where F_p^i are the membership functions of fuzzy sets, $i \in \{1, 2, ...r\}$, r is the number of rules, $A_i \in \mathbb{R}^{n*n}$, $B_i \in \mathbb{R}^{n*m}$, $C_i \in \mathbb{R}^{q*n}$ and $z_1(t), ..., z_p(t)$ are the premise variables which can be dependent of the input, the output or the state. The global T-S fuzzy model is given in the following form:

$$\begin{aligned} \dot{x}(t) &= \sum_{i=1}^{r} h_i(z(t))(A_i x(t) + B_i u(t)) \\ y(t) &= Cx(t) \end{aligned} \tag{3}$$

where

$$h_i(z(t)) = \frac{\prod_{j=1}^{p} F_j^i(z_j(t))}{\sum_{i=1}^{r} \prod_{j=1}^{p} F_j^i(z_j(t))} \tag{4}$$

The activation functions $h_i(z(t))$ indicates the activation degree of the ith associated local model, this functions verifies all time the convex sum propriety:

$$\begin{cases} 0 \le h_i(z(t)) \le 1 \\ \sum_{i=1}^{r} h_i(z(t)) = 1, \forall i \in \{1, 2, ..., r\}. \end{cases} \tag{5}$$

3 PDC control approach

3.1 Fuzzy regulator design via PDC

To stabilize the system presented by their T-S fuzzy model, the PDC controller is usually used to design a fuzzy controller. The main idea is to design a local controller for each sub-model based on local control rule, which shares with the fuzzy model the same fuzzy sets. The overall fuzzy controller is represented by:

$$u(t) = -\sum_{i=1}^{r} h_i(z) K_i x(t) \tag{6}$$

Where the K_i represent the local feedback gains. by using (6) in (3) the system in closed-loop becomes:

$$\begin{aligned} \dot{x}(t) &= \sum_{i=1}^{r} \sum_{j=1}^{r} h_i(z) h_j(z)(A_i - B_i K_j) x(t) \\ &= \sum_{i=1}^{r} h_i^2(z) G_{ii} x(t) + 2 \sum_{i<j}^{r} h_i(z) h_j(z) (\tfrac{G_{ij} + G_{ji}}{2}) x(t) \end{aligned} \tag{7}$$

with $G_{ij} = A_i - B_i K_j$, the stability conditions of (7) are given by the following theorem [17].

Theorem 1 *The continuous fuzzy system* (7) *is asymptotically stable, if there exist a common positive matrix* $P \in R^{n \times n}$ *and a common positive semi definite matrix* $Q \in R^{n \times n}$ *and for a number of active rules* s, *where* $1 < s \le r$ *such that:*

$$G_{ii}^T P + P G_{ii} + (s-1)Q < 0 \tag{8}$$

$$(\frac{G_{ij} + G_{ji}}{2})^T P + P(\frac{G_{ij} + G_{ji}}{2}) - Q \le 0, i < j \tag{9}$$

where $s > 1$.

In order to transform the preview conditions into LMIs form, one can consider the following variables: $X = P^{-1}$, $K_i = M_i X^{-1}$, $Q = PYP$, where $X > 0$, $Y \geq 0$ and $M_i (i = 1,...,r)$, then the stabilization conditions become:

$$\begin{aligned} &A_i X - B_i M_i + X A_i^T - M_i^T B_i^T + (s-1)Y < 0 \\ &A_i X - B_i M_j + X A_i^T - M_j^T B_i^T + A_j X - B_j M_i \\ &+ X A_j^T - M_i^T B_j^T - 2Y \leq 0, \qquad i < j \end{aligned} \tag{10}$$

4 Trajectory tracking control

The trajectory tracking control of nonlinear systems is the subject this section. In the tracking loop, we consider the integral of the tracking error $e_I = \int (y_r - y)dt = \int (y_r - Cx)dt$ [19], with y_r is the desired output. If we consider the following augmented state:

$$X_a = \begin{bmatrix} x \\ e_I \end{bmatrix}$$

Then, the following augmented system is obtained:

$$\begin{cases} \dot{X}_a(t) = \sum_{i=1}^{r} h_i(z(t))(\bar{A}_i X_a(t) + \bar{B}_i u(t) + \bar{D} Y_r) \\ y(t) = \bar{C} X_a(t) \end{cases} \tag{11}$$

where: $\bar{A}_i = \begin{bmatrix} A_i & 0 \\ -C & 0 \end{bmatrix}$, $\bar{B}_i = \begin{bmatrix} B_i \\ 0 \end{bmatrix}$, $\bar{C} = \begin{bmatrix} C & 0 \end{bmatrix}$, $\bar{D} = \begin{bmatrix} 0 \\ \mathbb{I} \end{bmatrix}$, $Y_r = \begin{bmatrix} y_{r1} \\ y_{r2} \end{bmatrix}$
with Y_r, denote the desired reference trajectory.

4.1 PDC control

To achieve the output tracking, the state feedback PDC control based on the previous LMIs can be used. The fuzzy controller $u(t)$ has the same form of (6), where x is replaced by the augmented state X_a:

$$u(t) = -\sum_{i=1}^{r} h_i(z) K_i X_a(t) = -\sum_{i=1}^{r} h_i(z)(K_{ix} x + K_{iI} e_I) \tag{12}$$

The feedback gains of the controller $K_i = \begin{bmatrix} K_{ix} & K_{iI} \end{bmatrix}$ are obtained by solving the LMIs (10).

4.2 LQI control

To design the LQI control, the following quadratic cost criterion must be minimized by the control law $u(t)$:

$$J = \int_0^{\infty} (X_a^T(t) Q X_a(t) + u^T(t) R u(t)) dt \tag{13}$$

for this reason, the following candidate quadratic Lyapunov function is considered:

$$V(X_a) = X_a^T P X_a \tag{14}$$

The augmented system (7) is stable if :

$$X_a^T Q X_a + u^T R u + \dot{V}(X_a) < 0 \tag{15}$$

$$(\bar{A}_i + \bar{B}_i K_i)^T P_i + P_i(\bar{A}_i + \bar{B}_i K_i) + Q + K_i^T R K_i < 0 \tag{16}$$

$$(\bar{A}_i X_i + \bar{B}_i Y_i)^T + (\bar{A}_i X_i + \bar{B}_i Y_i) + X_i Q X_i + Y_i^T R Y_i < 0 \tag{17}$$

$$(\bar{A}_i X_i + \bar{B}_i Y_i)^T + (\bar{A}_i X_i + \bar{B}_i Y_i) + Y_i^T R Y_i - X_i(-Q) X_i < 0 \tag{18}$$

by using the Schur complement procedure, the following stability conditions are obtained:

$$\begin{bmatrix} (\bar{A}_i X_i + \bar{B}_i Y_i)^T + (\bar{A}_i X_i + \bar{B}_i Y_i) & X_i & Y_i^T \\ * & -\tilde{Q} & 0 \\ * & * & -\tilde{R} \end{bmatrix} < 0 \tag{19}$$

where: $Q^{-1} = \tilde{Q}$, $R^{-1} = \tilde{R}$, $X_i = P_i^{-1}$, $K_i = Y_i X_i^{-1}$, $i = 1,...,r$
This LMIs will be calculated for each sub-model, here is about the conventional linear quadratic integral control using the fuzzy model. The control law is not based on fuzzy rules.

The control law will be respect the physical constraints on the control input, where this problem is studied in many practical cases [20, 21, 22, 23]. To ensure the stabilization under constraints on the inputs, the conditions given in the following theorem [17] must be verified.

Theorem 2 *For a known initial condition $X_a(0)$, the constraint $\|u(t)\|_2 \leq \eta$ is enforced at all times $t \geq 0$ if the LMIs:*

$$\begin{bmatrix} 1 & X_a(0)^T \\ X_a(0) & X \end{bmatrix} \geq 0 \tag{20}$$

$$\begin{bmatrix} X & M_i^T \\ M_i & \eta^2 I \end{bmatrix} \geq 0 \tag{21}$$

hold

4.3 Observer design

In bioprocess control problems, the state variables are not usually available. By introducing the observer, one can reconstruct partially or all the state variables. This section presents the fuzzy observer design with unmeasurable premise variables $z(t)$ ($h_i(z) \neq h_i(\hat{z})$).
Based on the structure of the fuzzy model (3), the fuzzy observer is given as follows:

$$\begin{cases} \dot{\hat{x}}(t) = \sum_{i=1}^{r} h_i(\hat{z})(A_i \hat{x}(t) + B_i u(t) + L_i(y(t) - \hat{y}(t)) \\ \hat{y}(t) = \sum_{i=1}^{r} h_i(\hat{z}) C_i \hat{x}(t) \end{cases} \tag{22}$$

where $\hat{x}$ denotes the estimated state and L_i the gains of the observer.
In order to compute this gains the following theorem [24] gives the necessary conditions for ensuring the convergence of the state estimation error to zero.

Theorem 3 *If there exist symmetric and definite positive matrices $P \in \mathbb{R}^{n \times n}$, $Q \in \mathbb{R}^{n \times n}$, matrices $Y_i \in \mathbb{R}^{n \times q}$ and a scalar $\alpha > 0$ such that:*

$$A_i^T P + P A_i - C^T Y_i^T - Y_i^T C \leq -Q \tag{23}$$

$$\begin{bmatrix} Q - \alpha^2 \mathbb{I} & P \\ P & \mathbb{I} \end{bmatrix} > 0 \tag{24}$$

then, the estimation error between the T-S fuzzy model (3) *and the fuzzy observer* (22) *is converges asymptotically to zero.*
where: $L_i = P^{-1}Y_i$.

Proof 1 *See Appendix.*

5 Process description

The proposed biological process in this paper is a biomass growth process, which consists to grow the population of microorganisms (biomass) by the consumption of a substrate (glucose), according to the following reaction scheme:

$$k_1 S \xmapsto{\mu(.)} X \tag{25}$$

The dynamic model of this process is established from the mass-balance [25], which describes the evolution of substrate and biomass concentrations in a continuous bioreactor. This model can be represented by a high nonlinear system as follows:

$$\begin{cases} \frac{dX}{dt} = \mu(.)X - DX \\ \frac{dS}{dt} = -k_1\mu(.)X + D(S_{in} - S) \end{cases} \tag{26}$$

The state variables are the biomass X and substrate S concentrations, k_1 denotes the pseudo stoichiometric coefficient and $\mu(.)$ represent the specific growth rate, the "Monod law" characterizes $\mu(.)$ is:

$$\mu(S) = \mu_{max}\frac{S}{K_s + S} \tag{27}$$

where μ_{max} is the maximum specific growth rate; K_s is the Monod or saturation constant. The input variables are the dilution rate $D(t)$ and the influent substrate concentration S_{in}. The parameters of the proposed model are given in the Table 1.

Parameters	Value	Unit
μ_{max}	0.38	h^{-1}
K_s	5	g/l
k_1	1/0.07	
S_{in}^{max}	140	g/l

Table 1: Simulation parameters

5.1 Takagi-Sugeno model design

The model (26) must be transformed into affine in control model like in (1), where the bioprocess models are known belong to the class of affine nonlinear models, this can be easily shown by assuming that:

$$\begin{aligned} D &= D_1 + D_2 \\ S_{in} &= \frac{D_2}{D_1+D_2}S_{in}^{max} \end{aligned} \tag{28}$$

where $D_1(t)$ and $D_2(t)$ are respectively the water and the substrate dilution rate, then one can replace $D(t)$ and $S_{in}(t)$ in (26) by their expressions (28), the following affine model is obtained:

$$\begin{bmatrix} \frac{dX}{dt} \\ \frac{dS}{dt} \end{bmatrix} = \begin{bmatrix} \mu(S)X \\ -k_1\mu(S)X \end{bmatrix} + \begin{bmatrix} -X & -X \\ -S & S_{in}^{max} - S \end{bmatrix}\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \tag{29}$$

where:

$$f(x) = \begin{bmatrix} \mu(S)X \\ -k_1\mu(S)X \end{bmatrix},\ g(x) = \begin{bmatrix} -X & -X \\ -S & S_{in}^{max} - S \end{bmatrix}$$
$$x(t) = \begin{bmatrix} X \\ S \end{bmatrix} \text{ and } u(t) = \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}$$

To build the T-S model, the following nonlinearities are considered:

$$z_1(x) = \mu(S) \tag{30}$$
$$z_2(x) = X \tag{31}$$
$$z_3(x) = S \tag{32}$$

This leads:
$A(z) = \begin{bmatrix} z_1 & 0 \\ -k_1 z_1 & 0 \end{bmatrix}$ and $B(z) = \begin{bmatrix} -z_2 & -z_2 \\ -z_3 & -z_3 + S_{in}^{max} \end{bmatrix}$
where the number of nonlinearities $n = 3$, the global model can be represented by $r = 2^n = 8$ sub-models. The local membership functions are defined by:

$$F_1^1(z_1) = \frac{z_1 - z_1^{min}}{z_1^{max} - z_1^{min}}, \quad F_1^2(z_1) = \frac{z_1^{max} - z_1}{z_1^{max} - z_1^{min}}$$
$$F_2^1(z_2) = \frac{z_2 - z_2^{min}}{z_2^{max} - z_2^{min}}, \quad F_2^2(z_2) = \frac{z_2^{max} - z_2}{z_2^{max} - z_2^{min}}$$
$$F_3^1(z_3) = \frac{z_3 - z_3^{min}}{z_3^{max} - z_3^{min}}, \quad F_3^2(z_3) = \frac{z_3^{max} - z_3}{z_3^{max} - z_3^{min}}$$

Finally, the activation functions are:

$$h_1(z) = F_1^1(z_1)F_2^1(z_2)F_3^1(z_3), \quad h_2(z) = F_1^1(z_1)F_2^1(z_2)F_3^2(z_3)$$
$$h_3(z) = F_1^1(z_1)F_2^2(z_2)F_3^1(z_3), \quad h_4(z) = F_1^1(z_1)F_2^2(z_2)F_3^2(z_3)$$
$$h_5(z) = F_1^2(z_1)F_2^1(z_2)F_3^1(z_3), \quad h_6(z) = F_1^1(z_1)F_2^1(z_2)F_3^2(z_3)$$
$$h_7(z) = F_1^2(z_1)F_2^2(z_2)F_3^1(z_3), \quad h_8(z) = F_1^2(z_1)F_2^2(z_2)F_3^2(z_3)$$

For the simulation, the parameters given in Table 1 are considered and leads to the following *min* and *max* of premise variables:

$$\begin{aligned} 0.018 &\le \mu(S) \le 0.35 \\ 3.8 &\le X \le 20 \\ 0.6 &\le S \le 140 \end{aligned}$$

the computed matrix A_i and B_i of each sub-model are given as follows:

$$A_1 = A_3 = A_5 = A_7 = \begin{bmatrix} 0.3507 & 0 \\ -5.0106 & 0 \end{bmatrix}$$
$$A_2 = A_4 = A_6 = A_8 = \begin{bmatrix} 0.0179 & 0 \\ -0.2551 & 0 \end{bmatrix}$$

$$B_1 = B_2 = \begin{bmatrix} -20 & -20 \\ -140 & 0 \end{bmatrix}, B_3 = B_4 = \begin{bmatrix} -3.8 & -3.8 \\ -140 & 0 \end{bmatrix}$$
$$B_5 = B_6 = \begin{bmatrix} -20 & -20 \\ -0.6 & 139.4 \end{bmatrix}, B_7 = B_8 = \begin{bmatrix} -3.8 & -3.8 \\ -0.60 & 139.4 \end{bmatrix}$$

6 Simulation and results

In the first case, all the states variables (substrate and biomass concentrations) are supposed measurable (i.e. $C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$).
The desired trajectory (X_r and S_r), which represent respectively of biomass and substrate concentrations are computed by using the following reference model:

$$\begin{aligned} \dot{X}_r &= -0.97X_r + 0.97ref_X \\ \dot{S}_r &= -0.65S_r + 0.65ref_S \end{aligned} \tag{33}$$

where ref_X and ref_S are the setpoints.

6.1 Tracking control based on state feedback

6.1.1 PDC control

The studied bioprocess presents the physical constraints on the control as shown in the Table 2

Variables	Constraints
Dilution rate h^{-1}	$0.01 \leqslant D \leqslant 0.38$
Influent substrate g/l	$60 \leqslant S_{in} \leqslant 140$

Table 2: The control constraints

For a number of active sub-model $s = 5$ and $\eta = 1.55$, the LMIs (10), (20) and (21) are solved by using the solver SeDuMi in MATLAB toolbox YALMIP, gives the following gains:

$$K_1 = \begin{bmatrix} 0.0384 & -0.0044 & -0.0007 & 0.0083 \\ -0.2821 & 0.0064 & 0.0564 & -0.0088 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 0.0052 & -0.0044 & -0.0007 & 0.0083 \\ -0.2472 & 0.0064 & 0.0563 & -0.0088 \end{bmatrix}$$

$$K_3 = \begin{bmatrix} 0.03470 & -0.0044 & 0.0002 & 0.0083 \\ -0.2963 & 0.0064 & 0.05930 & -0.0088 \end{bmatrix}$$

$$K_4 = \begin{bmatrix} 0.0011 & -0.0044 & 0.0002 & 0.0083 \\ -0.2605 & 0.0064 & 0.05910 & -0.0088 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -0.2149 & -0.0010 & 0.0572 & 0.0074 \\ -0.0287 & 0.0044 & -0.0014 & -0.0083 \end{bmatrix}$$

$$K_6 = \begin{bmatrix} -0.2480 & -0.0010 & 0.0571 & 0.0074 \\ 0.0060 & 0.0044 & -0.0015 & -0.0083 \end{bmatrix}$$

$$K_7 = \begin{bmatrix} -0.22450 & -0.0009 & 0.0592 & 0.0074 \\ -0.0371 & 0.0044 & 0.0003 & -0.0083 \end{bmatrix}$$

$$K_8 = \begin{bmatrix} -0.25660 & -0.0009 & 0.0590 & 0.0074 \\ -0.0028 & 0.0044 & 0.0003 & -0.0083 \end{bmatrix}$$

$$P = \begin{bmatrix} 0.0059 & -0.0000 & -0.0011 & 0.0000 \\ -0.0000 & 0.0008 & 0.0000 & -0.0002 \\ -0.0011 & 0.0000 & 0.0008 & -0.0000 \\ 0.0000 & -0.0002 & -0.0000 & 0.0008 \end{bmatrix}$$

$$Q = \begin{bmatrix} 7.5254 & 0.0002 & 2.9997 & -0.0011 \\ 0.0002 & 35.5015 & 0.0002 & 0.0002 \\ 2.9997 & 0.0002 & 23.9724 & 0.0013 \\ -0.0011 & 0.0002 & 0.0013 & 26.1931 \end{bmatrix}$$

The initial conditions are $x_0 = (6.6 \quad 5.50)^T$ and the obtained results are shown in Figures 1 and 2, the trajectory tracking is achieved, where the biomass and the substrate concentrations follow the desired outputs. In addition, the constraints on the inputs control $D(t)$ and $S_{in}(t)$ are respected.

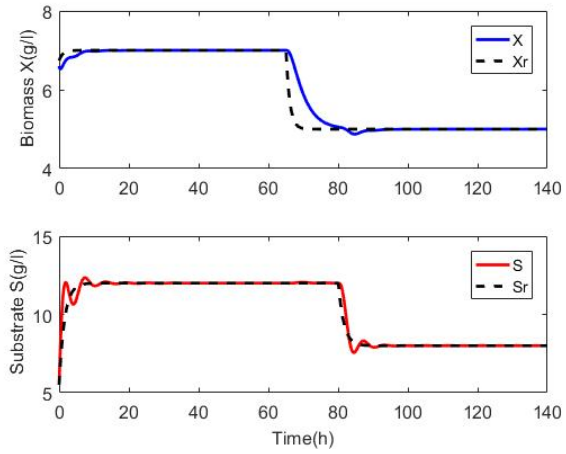


Figure 1: Evolution of the system outputs

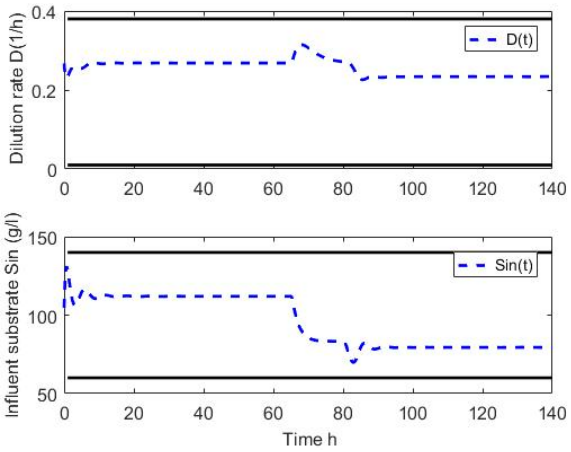


Figure 2: Control inputs $D(t)$ and $S_{in}(t)$

6.1.2 LQI control

Solving the LMIs established in (19), (20) and (21) the obtained weighting matrices are:

$$R = \begin{bmatrix} 0.328 & 0.000 \\ 0.000 & 0.328 \end{bmatrix}.10^{-3}$$

$$Q = \begin{bmatrix} 0.2529 & 0.0025 & -0.0264 & -0.0007 \\ 0.0025 & 0.2252 & -0.0004 & -0.0422 \\ -0.0264 & -0.0004 & 0.1345 & 0.0015 \\ -0.0007 & -0.0422 & 0.0015 & 0.1966 \end{bmatrix}.10^{-3}$$

The controller gains are:

$$K_1 = \begin{bmatrix} 0.0202 & -0.0081 & 0.0044 & 0.0101 \\ -0.0736 & 0.0128 & 0.0098 & -0.0056 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 0.0010 & -0.0095 & 0.0010 & 0.0105 \\ -0.0368 & 0.0052 & 0.0195 & -0.0030 \end{bmatrix}$$

$$K_3 = \begin{bmatrix} 0.0317 & -0.0095 & 0.0008 & 0.0106 \\ -0.2002 & 0.0039 & 0.0165 & -0.0004 \end{bmatrix}$$

$$K_4 = \begin{bmatrix} 0.0016 & -0.0096 & 0.0002 & 0.0106 \\ -0.1086 & 0.0023 & 0.0271 & -0.0005 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -0.0247 & -0.0003 & 0.0211 & 0.004 \\ -0.031 & 0.0095 & -0.0006 & -0.009 \end{bmatrix}$$

$$K_6 = \begin{bmatrix} -0.0345 & -0.0050 & 0.0200 & 0.0033 \\ -0.0024 & 0.0096 & 0.0009 & -0.0105 \end{bmatrix}$$

$$K_7 = \begin{bmatrix} -0.1471 & -0.0047 & 0.0211 & 0.0035 \\ -0.0369 & 0.0095 & 0.0003 & -0.0105 \end{bmatrix}$$

$$K_8 = \begin{bmatrix} -0.1064 & -0.0026 & 0.0274 & 0.0009 \\ -0.0023 & 0.0096 & 0.0003 & -0.0107 \end{bmatrix}$$

The Figures 3 and 4 show the simulation results using LQI control, where the controlled outputs variables achieve the desired outputs, also the constraints on control are respected.

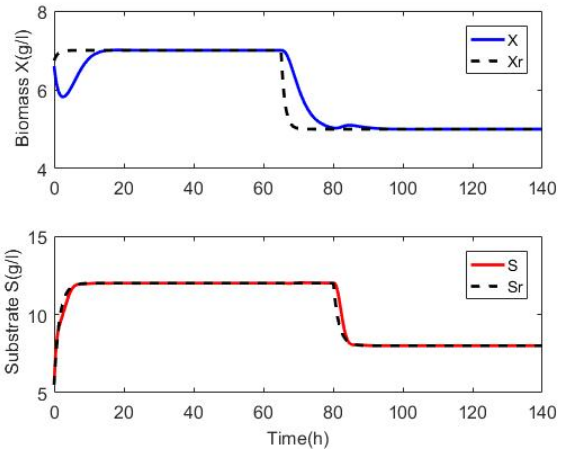


Figure 3: Evolution of the system outputs

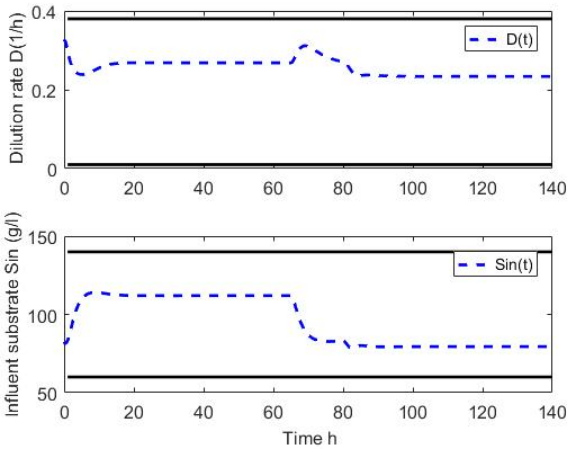


Figure 4: Control inputs $D(t)$ and $S_{in}(t)$

The results obtained are comparable or even better than those obtained using the PDC controller.

6.2 Tracking control based on reconstructed state feedback

In the practical case, only the substrate concentration is measured and the biomass is estimated, then the output matrix in (1) becomes $C = \begin{bmatrix} 0 & 1 \end{bmatrix}$. For the initial conditions of system $x_0 = (4 \quad 8)^T$, the initial conditions of observer $\hat{x}_0 = (5 \quad 7)^T$ and for a scalar $\alpha = 0.5$, the following observer gains are obtained:

$$L_1 = L_3 = L_5 = L_7 = \begin{bmatrix} 5.2458 \\ -1.3202 \end{bmatrix}$$

$$L_2 = L_4 = L_6 = L_8 = \begin{bmatrix} 0.3483 \\ -0.6440 \end{bmatrix}$$

$$Q = \begin{bmatrix} 5.655 & 0.000 & -0.000 \\ 0.000 & 4.4141 & 0.000 \\ -0.00 & 0.000 & 6.2045 \end{bmatrix}$$

$$P = \begin{bmatrix} 0.5985 & 0.0843 & 0.0000 \\ 0.0843 & 0.6103 & 0.0000 \\ 0.0000 & 0.0000 & 0.6103 \end{bmatrix}$$

6.2.1 PDC control

The resolution of the same LMIs in last case gives the following gains.

$$K_1 = \begin{bmatrix} 0.0231 & -0.0043 & 0.0067 \\ -0.1294 & 0.0058 & -0.0037 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} -0.0057 & -0.0044 & 0.0066 \\ -0.1082 & 0.0058 & -0.0033 \end{bmatrix}$$

$$K_3 = \begin{bmatrix} 0.0304 & -0.0044 & 0.0064 \\ -0.1425 & 0.0058 & -0.0027 \end{bmatrix}$$

$$K_4 = \begin{bmatrix} -0.0038 & -0.0045 & 0.0065 \\ -0.1187 & 0.0058 & -0.0031 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -0.0814 & -0.0015 & 0.0094 \\ -0.0336 & 0.0044 & -0.0065 \end{bmatrix}$$

$$K_6 = \begin{bmatrix} -0.1090 & -0.0015 & 0.0092 \\ -0.0016 & 0.0044 & -0.0064 \end{bmatrix}$$

$$K_7 = \begin{bmatrix} -0.0857 & -0.0018 & 0.0096 \\ -0.0359 & 0.0048 & -0.0065 \end{bmatrix}$$

$$K_8 = \begin{bmatrix} -0.1151 & -0.0015 & 0.0094 \\ -0.0036 & 0.0046 & -0.0064 \end{bmatrix}$$

$$P = \begin{bmatrix} 0.0114 & -0.0001 & -0.0003 \\ -0.0001 & 0.0010 & -0.0002 \\ -0.0003 & -0.0002 & 0.0008 \end{bmatrix}$$

$$Q = \begin{bmatrix} 0.0114 & -0.0001 & -0.0003 \\ -0.0001 & 0.0010 & -0.0002 \\ -0.0003 & -0.0002 & 0.0008 \end{bmatrix}$$

The Figure 5 shows a comparison between the controlled variable S, his estimated $\hat{S}$ and the desired output S_r. The obtained result show that S follows correctly S_r and the observer estimates the states of

system ($\hat{S} = S$) after 4.33 hours. The Figure 6 presents the control inputs $D(t)$ and $S_{in}(t)$, which respect the constraints.

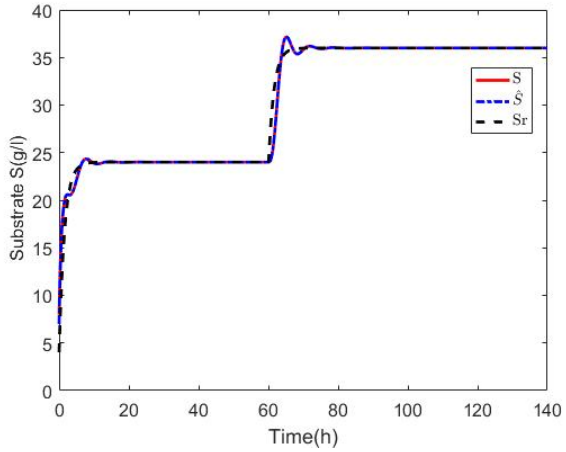


Figure 5: Evolution of the system outputs

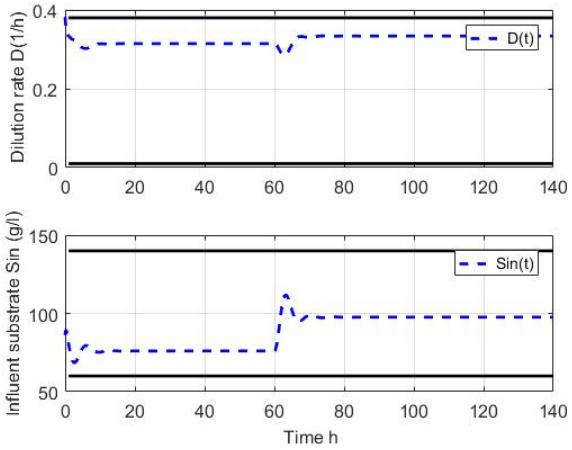


Figure 6: Control inputs $D(t)$ and $S_{in}(t)$

6.2.2 LQI control

In this part, we illustrate the obtained results using LQI control. Solving the LMIs (20), (21), (19) we obtain the following controller gains and weighting matrices:

$$R = \begin{bmatrix} 0.2040 & 0.0002 \\ 0.0002 & 0.2028 \end{bmatrix}.10^{-3}$$

$$Q = \begin{bmatrix} 0.1138 & 0.0012 & -0.0009 \\ 0.0012 & 0.1360 & -0.0196 \\ -0.0009 & -0.0196 & 0.1072 \end{bmatrix}.10^{-3}$$

$$K_1 = \begin{bmatrix} 0.0000 & -0.0043 & 0.0066 \\ -0.0566 & 0.0087 & -0.0035 \end{bmatrix}$$

$$K_2 = \begin{bmatrix} 0.0015 & -0.0076 & 0.0067 \\ -0.0124 & 0.0015 & -0.0001 \end{bmatrix}$$

$$K_3 = \begin{bmatrix} 0.0182 & -0.0072 & 0.0069 \\ -0.1308 & 0.0032 & -0.0003 \end{bmatrix}$$

$$K_4 = \begin{bmatrix} 0.0017 & -0.00760 & 0.0067 \\ -0.0222 & 0.0003 & 0.0002 \end{bmatrix}$$

$$K_5 = \begin{bmatrix} -0.0130 & 0.0005 & 0.0032 \\ -0.0231 & 0.0071 & -0.00560 \end{bmatrix}$$

$$K_6 = \begin{bmatrix} -0.0105 & -0.0017 & 0.0009 \\ -0.0016 & 0.0076 & -0.0067 \end{bmatrix}$$

$$K_7 = \begin{bmatrix} -0.0624 & -0.0034 & 0.0027 \\ -0.0327 & 0.0073 & -0.0062 \end{bmatrix}$$

$$K_8 = \begin{bmatrix} -0.0204 & -0.0009 & 0.0004 \\ -0.0018 & 0.0076 & -0.0067 \end{bmatrix}$$

The Figures 7 and 8 present the same variables in the Figures 5 and 6 using the LQI control. where the obtained results indicate clearly that the desired performances can be achieved more better than the results obtained by PDC control, the Figure 9 shows the state estimation error of the state variables (biomass and substrate concentrations), where this error tends to zero after 4.33 hours.

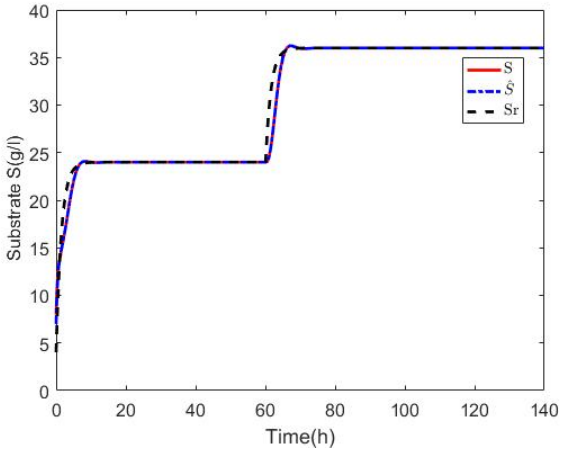


Figure 7: Evolution of the system outputs

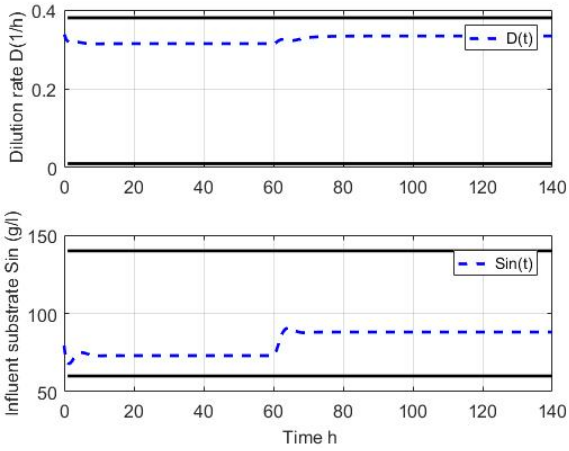


Figure 8: Control inputs $D(t)$ and $S_{in}(t)$

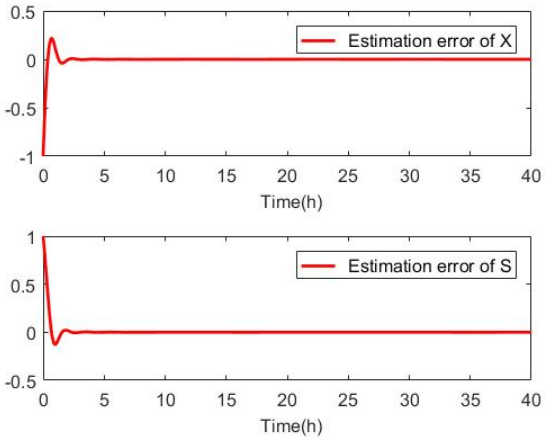


Figure 9: The sate estimation error

6.3 Comparison of both controllers

The outputs behaviors in the Figures 1, 3, 5, 7, show that the PDC controller presents some overshoot compared with LQI controller, where the PDC control requires the stabilization of all the different sub-models cross terms ($h_i(z) \neq h_j(z)$) too, which increases the LMIs need to be solved. Moreover, in terms of speed, the comparison shows clearly that the LQI controller is fast than PDC one. In general, the comparison shows that the LQI controller presents the best performance.

6.4 Comparison with other methods

Another method of bioprocess control based on neural network model has been developed in [13], where the process is controlled via a neural predictive controller in that case the controller doesn't refer to a mathematical model for the process but consider the black box identification. In our case we are based on the mathematical model for the synthesis of the controller, the obtained results are satisfactory regarding the result obtained in [13]. Another work in the same spirit developed by A. Nikfetrat [14], where the study considers a predictive controller applied to a biological feed-batch process without taking into account the constraints on the input variables and shows an important error in tracking reference trajectories, where the present work gave better results.

7 Conclusion

In this paper, the modeling and the control of the biomass growth process are treated. The objective is to control and compare two controllers. The nonlinear model of this process obtained from the mass-balance is transformed to Takagi-Sugeno fuzzy model, which represents exactly the original nonlinear model. Then, a T-S observer is designed to reconstruct the unmeasurable state when the premise variables are not measurable. To ensure the trajectory tracking two controllers are tested. The PDC and the linear quadratic control. The obtained results show that the two controllers are both effective. Finds that the second controller is more stable. In addition, the inputs respect the physical constraints of the process.
In this study, we only focus on the bioprocess control in faulty free case, that is why one interesting future work is to build a fault-tolerant control for this process.

Appendix

The proof of the Theorem 3 is presented here, we consider the following state estimation error:

$$e(t) = x(t) - \hat{x}(t) \tag{34}$$

their dynamic becomes:

$$\dot{e}(t) = \dot{x}(t) - \dot{\hat{x}}(t)$$

we take:

$$\triangle(x, \hat{x}, u) = \left(\sum_{i=1}^{r} h_i(z) - \sum_{i=1}^{r} h_i(\hat{z})\right)(A_i x + B_i u(t)) \tag{35}$$

the error dynamic then becomes:

$$\dot{e} = \sum_{i=1}^{r} h_i(\hat{z})(A_i - L_i C)e(t) + \triangle(x, \hat{x}, u) \tag{36}$$

if we assume that the term $\triangle(x, \hat{x}, u)$ satisfies the condition of Lipschitz as follows:

$$\| \triangle(x, \hat{x}, u)\| \leq \alpha \|x - \hat{x}\| \tag{37}$$

to ensure the convergence of (36) , one can consider the following candidate quadratic Lyapunov function:

$$V(e(t)) = e(t)^T P e(t) \tag{38}$$

leads to:

$$\dot{V}(e(t)) = \sum_{i=1}^{r} h_i(\hat{z}) e(t)^T ((A_i - L_i C)^T P + P(A_i - L_i C)) + \triangle(x, \hat{x}, u)^T P e(t) + e(t)^T P \triangle(x, \hat{x}, u) \tag{39}$$

Lemma 1 *For two real matrices X and Y of appropriate dimensions, the following inequality is verified:*

$$X^T Y + X Y^T < X^T \Omega^{-1} X + Y \Omega Y^T, \Omega > 0$$

Applying the previews lemma to the term:
$\triangle(x, \hat{x}, u)^T P e(t) + e(t)^T P \triangle(x, \hat{x}, u)$, the derivative of (39) is expressed as:

$$\dot{V}(e(t)) \leq e^T (A_i^T P - C^T L_i^T P + P A_i - P L_i C + \alpha^2 \mathbb{I} + P P) e \tag{40}$$

If there exists a symmetric and positive definite matrix $Q = Q^T$ such that:

$$A_i^T P - C^T L_i^T P + P A_i - P L_i C \leq -Q \tag{41}$$

leads to:

$$-Q+\alpha^2\mathbb{I}+PP<0 \tag{42}$$

Finally, we apply the Schur complement to (42) we obtain:

$$\begin{bmatrix} Q-\alpha^2\mathbb{I} & P \\ P & \mathbb{I} \end{bmatrix}>0 \tag{43}$$

Then the proof is completed.

References

[1] V. Vastemans, M. Rooman, P. Bogaerts, "A robust method for the joint estimation of yield coeffcients and kinetic parameters in bioprocess models", Biotechnology Progress 25 (3), 2009, 606–618. doi:10.1002/btpr.89.

[2] A. Karama, O. Bernard, J.-L. Gouzé," Constrained Hybrid Neural Mod- elling of Biotechnological Processes", International Journal of Chemical Re- 70 actor Engineering 8 (1), 2010, 1–15. doi:https://doi.org/10.2202/ 1542-6580.2117.

[3] V. Grisales, A. Gauthier, G. Roux, "Fuzzy Optimal Control Design for Dis- crete Affine Takagi-Sugeno Fuzzy Models: Application to a Biotechnologi- cal Process", 2006 IEEE International Conference on Fuzzy Systems, 2006, 2369–2376, doi:10.1109/FUZZY.2006.1682030.

[4] E. Herrera, B. Castillo, J. Ramrez, E. C. Ferreira, "Takagi-sugeno multiple- model controller for a continuous baking yeast fermentation process", ICINCO 2007 - 4th International Conference on Informatics in Control, Automation and Robotics, Proceedings ICSO, 2007, 436–439. doi:10.5220/0001622704360439.

[5] Min-Sen Chiu, Shan Cui, Qing-Guo Wang. " Internal Model Control Design for Transition Control", AIChE journal, vol 46, 2000

[6] E. J. Herrera-Liopez, B. Castillo-Toledo, R. Femat, "Fuzzy servo controller 65 for CSTB with substrate inhibition kinetics", Journal of Process Control 22 (6) (2012) 959–967. doi:10.1016/j.jprocont.2012.05.003.

[7] A. M. Nagy Kiss, B. Marx, G. Mourot, G. Schutz, J. Ragot, "Observers de- sign for uncertain Takagi-Sugeno systems with unmeasurable premise vari- ables and unknown inputs. Application to a wastewater treatment plant", 30 Journal of Process Control 21 (7), 2011, 1105–1114. doi:10.1016/j. jprocont.2011.05.001.

[8] M. Bouharkat, M. Ramdani," Fuzzy observer based predictive control of an activated sludge depollution bioprocess", 2013 International Conference on Control, Decision and Information Technologies, CoDIT, 2013, 236–241. doi:10.1109/CoDIT.2013.6689550.

[9] E. Herrera, B. Castilloa, J. Ramrezb, E. C. Ferreirac, "Exact fuzzy observer for a baker's yeast fermentation process", IFAC Proceedings Volumes (IFAC- PapersOnline) 10 (1), 2007, 313–318. doi:10.1109/FUZZY.2007.4295502.

[10] S. Aouaouda, M. Chadli, M. T. Khadir, T. Bouarar, "Robust fault tolerant tracking controller design for unknown inputs T-S models with unmeasurable premise variables", Journal of Process Control 22 (5), 2012, 861–872. doi:10.1016/j.jprocont.2012.02.016.

[11] S. Bououden, M. Chadli, H. R. Karimi, "Control of uncertain highly non- linear biological process based on Takagi-Sugeno fuzzy models", Signal Processing 108 (2015) 195–205. doi:10.1016/j.sigpro.2014.09.011.

[12] S. Carlos-Hernandez, E. N. Sanchez, J. F. Béteau, "Fuzzy observers for anaerobic WWTP: Development and implementation", Control Engineering Practice 17 (6), 2009, 690–702. doi:10.1016/j.conengprac.2008.11. 008.

[13] A. Karama, "Contribution à la modélisation et la commande par réseaux de neurones des bioprocédès", thesis, University Cadi Ayyad, Faculty of Science Semlalia Marrakech, 2004.

[14] A. Nikfetrat, A.R. Vali, V. Babaeipour. "Neural Network Modeling and Nonlinear Predictive Control of a Biotechnological Fed-batch Process", IEEE International Conference on Control and Automation Christchurch, New Zealand, December 9–11, 2009.

[15] Mohd N, Aziz N. "Performance and Robustness Evaluation of Nonlinear Autoregressive with Exogenous Input Model Predictive Control in Controlling Industrial Fermentation Process", Journal of Cleaner Production, 2016.

[16] O. Gehan, E. Pigeon, M. Pouliquen, L. Fall, R. Mosrati, "Nonlinear control of dissolved oxygen level for Pseudomonas putida bacterium fermentation", 15 2016 IEEE Conference on Control Applications, CCA 2016 (2016) 1215– 1220 doi:10.1109/CCA.2016.7587972.

[17] K. Tanaka, H. O. Wang, "Fuzzy Control Systems Design and Analysis", John Wiley and Sons, Inc., New York, USA, 2001. doi:10.1002/0471224596.

[18] Saifia, D., Chadli, M., Labiod, S., and Guerra, T. M., " Robust $H\infty$ Static Output Feedback Stabilization of T-S Fuzzy Systems Subject to Actuator Saturation", Int. J. Control Autom. Syst. 10(3), 613–622, 2012

[19] Hassan.K.Khali. "Nonlinear Systems", second edition, Prentice Hall. United States of America, 1996.

[20] Guechi, El-Hadi, Lauber, Jimmy, Dambrine, Michel, et al., "PDC control design for non-holonomic wheeled mobile robots with delayed outputs", J. Intell. Robot. Syst. 60(3-4), 395–414, 2010

[21] Xu, Bin, Wang, Shixing, Gao, Daoxiang, et al., "Command filter based robust nonlinear control of hypersonic aircraft with magnitude constraints on states and actuators", J. Intell. Robot. Syst. 73(1-4), 233–247, 2014

[22] Zhu, Senqiang et Wang, Danwei., " Adversarial ground target tracking using UAVs with input constraints", J. Intell. Robot. Syst. 65(1-4), 521-532, 2012

[23] M. Abyad, A. Karama, A. Khalloq, "Fuzzy Takagi-Sugeno Based Modelling and Control For an Alcoholic Fermentation Process", 3rd International Conference on Electrical and Information Technologies ICEIT, Rabat, Morocco, 2017.

[24] P. Bergsten, R. Palm, D. Driankov, "Fuzzy observers", 10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297) 2, 2001, doi:10.1109/FUZZ.2001.1009051.

[25] Bastin G. and Dochain D. "On-line estimation and adaptive control of bioreactors". Elsevier, Amsterdam, 1990.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 319-323 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Metaheuristics for Solving Facility Location Optimization Problem in Lagos, Nigeria

Chika Yinka-Banjo*, Babatunde Opesemowo

Computer Science, University of Lagos, Akoka, Lagos, 100213, Nigeria

ARTICLE INFO

Article history:
Received: 24 August, 2018
Accepted: 10 November, 2018
Online: 25 November, 2018

Keywords:
Metaheuristics
Facility Location Problem
Optimization

ABSTRACT

Facility location problem is a problem that many organizations still face today because of its increasing constraints and objectives. Decision makers want this problem solved in order to maximize profit and as such, it became a field of interest to many computer scientists over the years. The solution tool used by these scientists; a function of technological advancement, has evolved from the use of classical mathematical approaches to the use of metaheuristics. Some of the metaheuristics used include particle swarm optimization metaheuristics, genetic algorithm metaheuristics and tabu search. The problem considered in this research evolves the study of waste management in Lagos state. How the location of waste evacuation centers could be allocated in order to minimize resources such as transportation cost, facility cost, distance and the capacity of each centers. A mathematical model was developed that serves as a template for the algorithm used to solve the problem. Then particle swarm optimization metaheuristics was used to find the optimal solution in terms of capacity to the problem. Particle swarm optimization minimizes the use of memory and still gives a satisfactory solution. With the result obtained, respective agencies could make good decision as to the best location to build a new facility.

1. Introduction

The study of facility location problem, also known as location analysis, is a branch of operations research and computational geometry concerned with the optimal placement of facilities to minimize resources such as cost, time, distance, etc. In most cases, the cost minimized is transportation cost. There are different areas facility location problem could be considered. These areas may include logistics facilities, warehouses, farm facilities, police stations etc. For this project, we considered a waste management facility known as evacuation centres. Waste generated from the residents at the state are dumped at the evacuation centres where it is recycled. We chose the waste management facility problem because of the way it affect our environment especially the Lagos state government. The state is very keen to the management of waste generated because of her thriving effort to becoming a green state. In this effort, the government has employed private own company known as Visionscape for the management of her waste. We have approximately 21 million people living in Lagos state and quite a large amount of waste are generated in the state daily. When the waste are not well managed based on poor allocation of the evacuation centres this could lead to air pollution of the environment. However, selecting locations out of the blues are not best measures to obtain a reasonable result. Hence, after the study of major works done on facility location, we embark on finding scientific ways of solving the problem that will give a satisfactory result. This led to the use of metaheuristics.

Metaheuristics is a higher-level procedure or heuristic designed to find, generate, or select a heuristic (partial search algorithm) that may provide a sufficiently good solution to an optimization problem, especially with incomplete or imperfect information or limited computation capacity. Examples of metaheuristics are particle swarm optimization (PSO), hill climbing, simulated annealing, tabu search, genetic algorithm, etc. Figure 1 shows the classification of metaheuristics. Different analysis have been used over the years for these classes to be obtained and presented as a clear cut direction [1].

2. Related Literature

Over the past decades, the study of facility location problem has grown even more popular, not only in the academic literature but also in practice. Facility location problems are often strategic in nature and entail long-term decisions, exposing firms to many uncertainties during the operational lifetime of a facility. When solving such problems, a firm may have to determine the number of facilities to open, their locations, and their capacities. Because

*Chika Yinka-Banjo, Email: cyinkabanjo@unilag.edu.ng

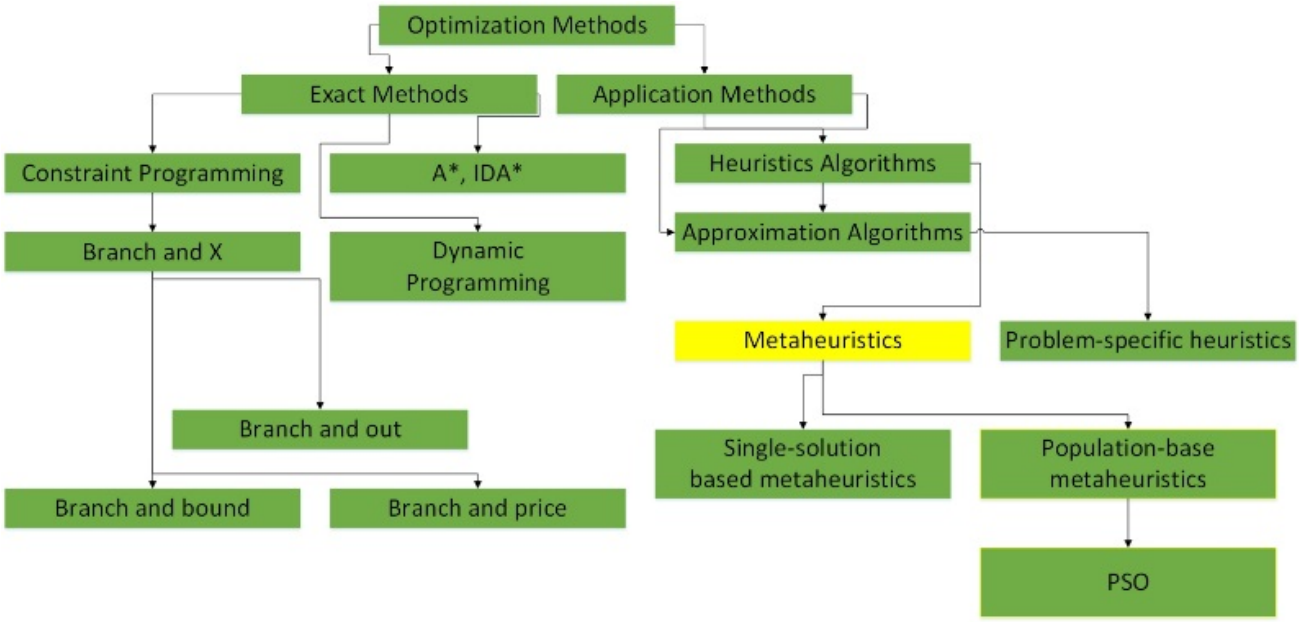


Figure 1: Classification of metaheuristics.

of the high fixed costs incurred in changing a network of facilities, a firm may be limited in the frequency in which it re-examines these strategic decisions. After determining its facility network, the firm is often relegated to making operational decisions such as determining production quantities, service levels, and allocation of supply to demand. Thus, facility location problems often have a two-stage approach to solving them, location followed by evaluation. Facility location problem belongs to the class of a combinatorial optimization problem [2].

The 17th Century marks the beginning of the facility location problem which was considered as a classical mathematical problem [3]. As time went on, it became an interest to many scientists and researchers because of how it relate to real life issues. Also, various aspect of facility location problem have been considered by earlier scientist ranging from customer facility management to logistics management. [4]. However, the problems are majored on customer and warehouse facility. Also, facility location problem is further considered as capacitated [5, 6, 7] and uncapacitated [8]. Capacitated in terms of considering the size of the facility involved or the ability of the facility to manage other entities involved like customer adequately. While uncapacitated doesn't consider the size of the facility, it assumes the facility could always manage the entities involved. Both capacitated and uncapacitated were considered by Bumb and March [9].

Mauricio and Renato presented a multistart heuristic for the uncapacitated facility location problem [10]. The method used combines elements of several other metaheuristics, such as scatter and tabu search (which make heavy use of path-relinking) and genetic algorithms (that deals with the notion of generations) to proffer solutions.

In order to have a good understanding of the various state of facility location problems, Andreas et al. [11] studied continuous location models, network location models, mixed-integer programming models, and applications of facility location problem. In [12], the author proposed particle swarm optimization, simulated annealing and iterated local search metaheuristics for solving facility location problem that has to do with the health sector. The problem was modelled and the experimental results show that the proposed algorithms reach acceptable performances in a reasonable computation time. In [13], the auhtors considered forty hospitals and three candidate municipalities in the sub-northeast region of Thailand, and considered multiple factors such as infrastructure, geological and social & environmental factors, and calculating global priority weights using the fuzzy analytical hierarchy process (FAHP). Also, a new multi-objective facility

location problem model which combines FAHP and goal programming (GP), namely the FAHP-GP model, was tested. Their proposed model led to selecting new suitable locations for infectious waste disposal by considering both total cost and final priority weight objectives. The novelty of the proposed model is the simultaneous combination of relevant factors that are difficult to interpret and cost factors which require the allocation of resources.

From the above literature, facility location problems cut across different horizon of research, and few research have provided an overview of the models and algorithms that are applied to the optimization in solving facility location problem that has to do with waste management.

3. Methodology

In this paper, we studied the problem of allocation of new evacuation centres at different potential locations such that it could manage the local government areas assigned to it in relation to the existing potential locations. The major entities are evacuation centres, potential location and local government areas. Evacuation centres are spaces where wastes are dumped and recycled. Whereas, potential locations are the positions where a facility can be allocated. Finally, the waste are generated from the local government areas in a particular city. The problem was well formulated following the principles that governs optimization model. The objective of the optimization model is to minimize the total cost of evacuation sites to be located, chosen among a set of

candidate locations. Such objective ensures not only the reduction of the visual impact due to the presence of collection sites, but also the reduction of the overall cost related to the collection phase. Following below are the indices, parameters and decision variables for the model.

3.1. Definition of model components

INDICES:

i = is the index of each potential location(Town) in the state, i = 1,2,…m (m=3)

j = is the index of each Evacuation Center, j = 1,2,…n (n=2)

k = is the index of each Local Government Area (LGA), k = 1,2,…t (t=21)

PARAMETERS:

f_i = is the facility cost (Naira)

w_i = is the weight of the factor affecting a location

p_{ik} = is the transportation cost between the potential location and LGA(Naira)

d_{ik} = is the distance between the potential location and LGA(km)

s_j = is the size of each evacuation center

x_k = is the quantity of refuse generated by a LGA

DECISION VARIABLES:

Y_{ik} = is a binary variable where it is 1 if the LGA is served by Potential Location k and 0 if not served

H_i = is a positive integer value where it is 1 if the Potential Location is selected and 0 if not selected

O_{ij} = is a binary variable where it is 1 if the Potential Location is opened by selecting Evacuation Center j and 0 if otherwise

$$min.Z = \sum_{i=1}^{m}\sum_{j=1}^{n} f_i O_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n} w_{ij} O_{ij} + \sum_{i=1}^{m}\sum_{k=1}^{t} p_{ik} d_{ik} Y_{ik} \quad (1)$$

Subject to:

$$\sum_{i=1}^{m} Y_{ik} = 1 \qquad \forall k \{1,2,\dots t\} \quad (2)$$

$$\sum_{j=1}^{n} s_j O_{ij} \geq \sum_{k=1}^{t} x_k Y_{ik} \qquad \forall i \{1,2,\dots m\} \quad (3)$$

$$\sum_{j=1}^{n} O_{ij} = H_i \quad (4)$$

$$O_{ij} \in \{0,1\} \quad (5)$$

$$Y_{ik} \in \{0,1\} \quad (6)$$

$$H_i \in \{0,1\} \quad (7)$$

Equation (1) is the objective function which is defined to minimize the total cost based on the facility cost, factors affecting the potential location and the transportation cost. Equation (2) ensures that the demand of each LGA k is met. Equation (3) ensures that the sum of service by the LGA does not exceed the capacity of the Evacuation Center. Equation (4) ensures that the selected LGAs must use k-size Evacuation Centers. Constraints (5), (6) and (7) are binary decision variables. From this model, we applied particle swarm optimization technique to proffer an optimized solution on the capacity for each evacuation centers. For the PSO algorithm the following parameters below were used in relation to what was used by Jordanski:

- Initial weight w = 0.9
- Number of particles = 40
- Cognitive constant c1 = 2
- Social constant c2 = 2
- Maximum velocity vmax = capacity_range[0]
- Minimum velocity vmin = capacity_range[1]

Pseudocode 1: To find the optimal location for the evacuation centers

- Initialize Parameters, Decision Variables and Indices
- For (All number of Locations) {
  For (All number of LGA) {
  ObjCost ← FacilityCost + WeightCost + TransportationCost
  }
  ObjCostRow [ ] ← ObjCost
- }
- MinObjCost ← sort(ObjCostRow)
- MinCapacity ← Pso(Capacity)
- Return MinObjCost, MinCapacity

Pseudocode 1 shows the steps involved to solve the facility location problem for the evacuation center. All the parameters, decision variables and indices where initialized. An iteration was done to get the mapping for the potential location to their respective local government area (LGA). This will enable us generate the network analysis for the problem and find the minimum mapping that will guarantee a minimum cost based on all the criteria considered. The single cost for each mapping are placed in an array called ObjCostRow as shown above. It is sorted in an ascending order and pass to MinObjCost. The capacity generated for each evacuation center is passed to the PSO function which returns the minimal capacity for evacuation centers. Lastly, the result are displayed.

4. Application of the proposed methodology

The method was implemented with the PHP programming language and designed with web technologies such as HTML and CSS to make it easy for users to interact with. All the user needs to do is to supply the required parameters needed in the right form and the system validates the parameters provided and process it. Finally, the result is displayed to the user.

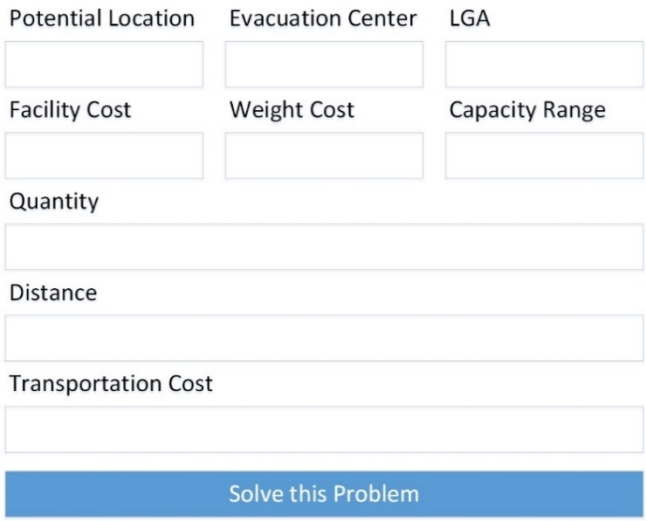


Figure 2: User interface for the parameters.

- **Potential Location:** This field requires the user to enter the number of potential location in consideration. The potential location is the location where the user desire to locate the facility. The value must be an integer.
- **Evacuation Centre:** This field requires the user to enter the number of evacuation center required to be allocated at the potential location. The value must be an integer.
- **LGA:** This field requires the user to enter the Local Government Area in the state where the refuse are coming that are to be managed by the evacuation center. The value must be an integer.
- **Facility Cost:** This field requires the user to enter the cost of building as the case may be at a particular potential location. The values are separated by a comma. The values must be a set of numbers and must correspond to the number of potential location set.
- **Weight Cost:** This field requires the user to enter the cost of the weight generated for each potential location. The weight is generated by considering the factors affecting each potential location. This factors may include geographical factor, closeness to raw materials, electricity, good roads, etc. The values are separated by a comma. Also, the values must correspond to the number of potential location set.
- **Capacity Range:** This field requires the user to enter the capacity search space for the problem. The capacity that is required to be managed for each evacuation center. This allows the user to know the required capacity to be considered for building a particular facility.
- **Quantity:** This field requires the user to enter the amount of waste in kilograms (Kg) generated by a particular LGA per day/month/year. The values could be a set of real numbers and separated by commas.
- **Distance:** This field requires the user to enter the distance from a particular LGA to a potential location in miles. The distance of the set of LGA is separated by commas and separated by a forward slash with respect to a particular potential location.
- **Transportation Cost:** This field requires the user to enter the transportation cost for each LGA to a particular potential location. The set of transportation cost are separated by commas and separated by a forward slash with respect to a particular potential.

We carried out 3 different experiment base on the number of potential locations, evacuation centers and local government areas. We shall be considering the operation of experiment one and give a summary of the three experiment.

Table 1: Parameters for Experiment 1

m	n	t	f_i	w_i	s_j
3	2	21	10,20,30	50,70,40	20,0
x_k	4,3,4,5,2,3,3,4,6,5,6,3,4,5,2,3,3,4,6,5,6				
d_{ik}	2,1,2,3,4,5,6,6,8,8,9,9,1,2,1,2,3,3,3,3,2/ 3,2,3,4,5,1,2,2,9,2,2,3,4,5,2,3,3,4,6,5,6/ 4,3,4,5,2,3,3,4,6,5,6,3,4,5,2,3,3,4,6,5,6				
p_{ik}	2,1,2,3,4,5,6,6,8,8,9,9,1,2,1,2,3,3,3,3,2/ 3,2,3,4,5,1,2,2,9,2,2,3,4,5,2,3,3,4,6,5,6/ 4,3,4,5,2,3,3,4,6,5,6,3,4,5,2,3,3,4,6,5,6				

Table 1 shows that we have 3 potential locations but we want to build only 2 evacuation centers that will minimize general cost (facility cost, transportation cost, and weight cost). Using the arbitrary parameters from Table 1, the objective function was generated following the operations described in Pseudocode 1 is shown in Figure 3.

======OBJECTIVE FUNCTION====

OPTIONS/FACILITY	COST	A	B	C
1	967	Close	Open	Open
2	987	Open	Close	Open
3	892	Open	Open	Close

Figure 3: Objective function result for experiment one.

From the objective function shown in Figure 3, we would see that the minimum cost is given by option 3. This option means the best selection would be facility A and B. The network analysis that gives this cost is shown in Figure 4. Hence, facility A would serve LGA (1,2,3,4,5,9, 13,14,15,16,17,18,19,20,21) while facility B serve LGA (6,7,8,10, 11,12).

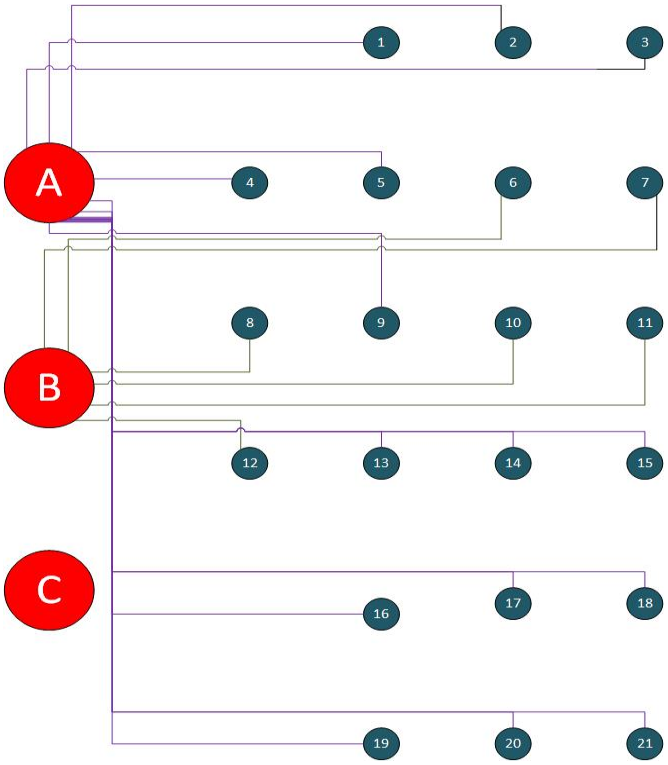


Figure 4: Network analysis for experiment one.

======CAPACITY============

FACILITY	CAPACITY
A	64
B	12
C	0

Figure 5: Capacity for experiment one.

Figure 5 shows the capacity for the facilities. That is, the minimum quantity a facility should manage. The same process as explained in Section 4 was carried out for two different scenerios with respect to the potential location, evacuation center and LGA. The time duration was recorded to evaluate the scalabity of the solution in relation with time as shown in Figure 6 and 7.

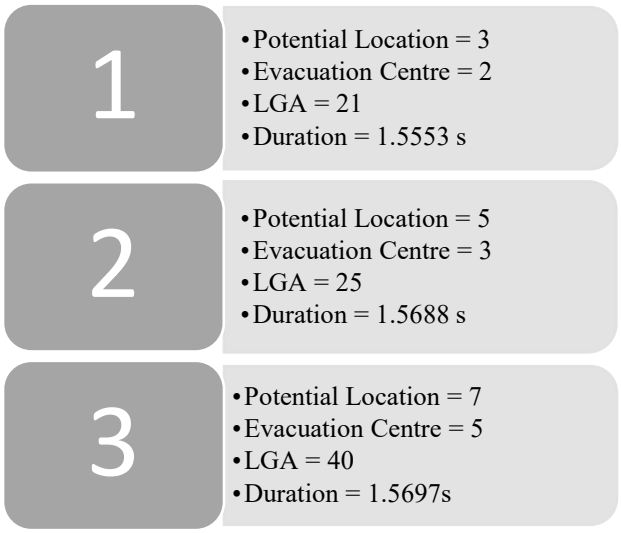


Figure 6: Summary of the experiments.

From Figure 7, we could discover that the time duration compare to the other parameters is significantly small.

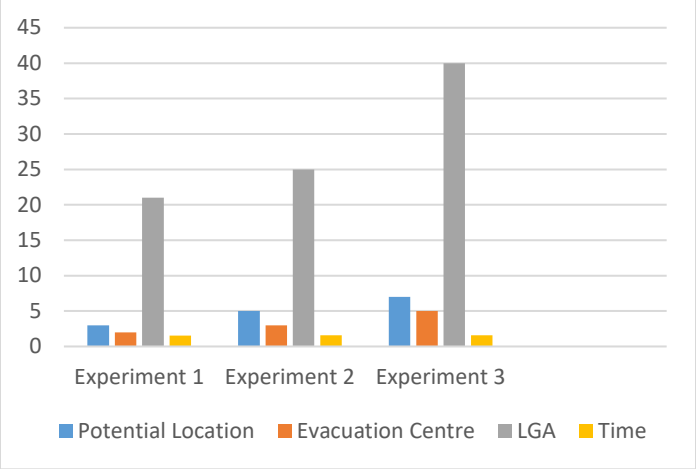


Figure 7: Graphical representation of the experiments.

5. Conclusion

In this project, facility location problem was considered in regards to optimization. The waste management facility serve as the underlying problem dealt with among several problems related to the facility location problem. This problem was chosen because of it adverse effect in the economy of Nigeria especially Lagos state. A redefined PSO was proposed in solving the problem which involves finding the best location to build a set of evacuations canters that would manage the waste gotten from several local government. As most common with facility location problem, the mathematical model to the problem was given which includes the objective functions and the set of constrains involved. Then a set of instances was used to test the model. The experimental result was gotten that shows the optimal solution to the problem. Hence, we have been able to develop a model that can be used to solve the facility location problem and proffer an optimized solution to the waste management facility location problem. The application developed can also accept other values which requires solution relating to the model worked on making it a flexible solution. Further research should be carried out in other problem areas like the police station facility location, farm facility allocation and any of those that could help in building the nation. Metaheuristics are very good tool in regards to problems that tends to be complex and are unable to be solve within a limited amount of time. However, they might not guarantee optimal solution in some cases. Hence, they must be applied carefully to solve a particular problem.

References

[1] E. G. Talbi, Metaheuristics - From Design to Implementation (John Wiley & Sons, New Jersey, 2009)
[2] S. Arifin. Location allocation problem using genetic algorithm and simulated annealing. A case study based on school in Enschede (Doctoral dissertation, Master's thesis, the University of Twente, Enschede, the Netherlands, 92p) (2011)
[3] R. Guner and S. Mehmet. A Discrete Particle Swarm Optimization Algorithm for Uncapacitated Facility Location Problem. Journal of Artificial Evolution and Applications, (2008).
[4] R.H Ballau, Facility Location Decisions. In Business Logistics: Supply Chain Management (2004) 550-617
[5] Ling-YunWua, Xiang-Sun Zhang, Ju-Liang Zhang. Capacitated facility location problem with general setup cost. Computers & Operations Research, (2006).
[6] H. Pirkul and V. Jayaraman. A Multi-Commodity, Multii-Plant, Capacitated Facility Location Problem: Formulation and Efficient Heuristic Solution. Elsevier Science Ltd, (1997).
[7] Y. Ren. Metaheuristics for multiobjective capacitated location allocation on logistics networks (Doctoral dissertation, Concordia University, 2011).
[8] R. L. Francis, L. F. McGinnis and J. A. White, Facility Layout and location: an analytical approach (Pearson College Division, 1992).
[9] A. Bumb and W. K. March. A simple dual ascent algorithm for the multilevel facility location problem: Approximation, Randomization, and Combinatorial Optimization. University of Twente (2001) 55-63
[10] G. R. Mauricio and F. W. Renato. A hybrid multistart heuristic for the uncapacitated facility location problem. European Journal of Operational Research, (2005).
[11] A. Klose and A. Drexl. Facility location models for distribution system design. European Journal of Operational Research, 162 (2005) 4-29.
[12] M. Jordanski. Metaheuristic Approaches for Solving Facility Location and Scale Decision Problem with Customer Preference. IPSI Bgd Internet Research Society, 13 (2017).
[13] N. Wichapa, P. Khokhajaikiat. Solving multi-objective facility location problem using the fuzzy analytical hierarchy process and goal programming: a case study on infectious waste disposal centers. Elsevier Science Ltd, (2017).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 321-327 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

The Value of Integrating MSRP Protocol in E-learning Platforms of Universities

Khalifa Sylla[1,*], Samuel Ouya[2], Masamba Seck[3], Gervais Mendy[2]

[1] *Virtual University of Senegal, Applied Mathematics and Computer, 15126 Dakar-Fann, Senegal*

[2] *Polytechnic High School of Dakar, Computer Science Department, 5085 Dakar-Fann, Senegal*

[3] *Alioune Diop University of Bambey, Computer Science Department, 16 617 Dakar Fann, Senegal*

ARTICLE INFO

Article history:
Received: 18 August, 2018
Accepted: 01 October, 2018
Online: 14 October, 2018

Keywords:
E-learning
STEM evaluation
Learning Management System
Virtual classroom
MSRP
WebSocket

ABSTRACT

The proposed solution is intended for virtual universities wishing to improve their distance learning platforms. These enhancements include the instructional content model, how to deliver courses to students, and how to evaluate students' knowledge. With this solution that improves the functionality of Moodle, virtual classes now include new features such as the transfer of heavy files, control of remote screens, video, audio, chat without degradation of the quality of service. The solution will be particularly useful for language teachers who have difficulty in assessing the speaking skills of their students. Indeed, the platform with its audio / video features, integrates a tool to assess the knowledge and improve the speaking skills of students. Teachers of technical subjects such as programming or computer networks will also be able to use this solution to improve the way they deliver courses, including making it easier for students to do practical work.
On the technical side, the solution is based on the MSRP-RELAY, MSRP and Websocket protocols. It is therefore compatible with smartphones, tablets and computers. Tests were conducted with teachers of languages and STEM subjects to evaluate their students in direct synchronous encounters without necessarily going through multiple choice questions.

1. Introduction

This paper is an extension of work originally presented in conference International Conference of Advanced

Communications Technology (ICACT) 2018 in Korea [1]. Currently, in many countries, particularly in Africa, higher education faces several constraints, including the lack of facilities to accommodate new graduates in traditional universities and teaching materials for practical work.

With the democratization of the Internet and the emergence of new ICT services, we can say that technology is at the service of education. Many countries are establishing national interconnection of public buildings by fiber optics. This promotes high bandwidth internet access, the ability to do cloud computing, and so on. Classical universities have taken advantage of these services and now offer Open and Distance Training [2]. For a purely virtual university, it is therefore a challenge to offer solid platforms with advanced features that allow students to carry out practical work while guaranteeing quality in audio-video communications during these practical work, especially for students. students in STEM [3].

Indeed, open and distance learning was once essential for professionals seeking skills development. Today, with digital universities, a young graduate from the bachelor's degree can start his university course with distance education. This explains why most Senegalese universities are increasingly interested in this educational innovation in their training offerings.

It is therefore very important for these young people to receive quality training, especially those who study STEM. Among other features of the training platforms used by universities to meet this need include communication tools (chat, audio, video), online self-assessment tools, [4] and so on. The Moodle platform [5] is undoubtedly one of the most used today by universities.

Moodle's very popular BigBlueButton [6] plugin allows you to create virtual classes using the WEBRTC [7] protocol. It includes all the tools you need for a virtual classroom (audio, video, screen sharing using flash [8]), but there are some shortcomings that

*Corresponding Author: Khalifa Sylla, Email: khalifa.sylla@uvs.edu.sn

deserve to be raised. The possibility of entering mathematical formulas and the impossibility of taking remote control of computers seriously handicap the dispensing of certain courses and, by extension, the quality of the students' training.

Given the requirements of distance learning today, we believe that an e-learning platform should no longer be limited to the traditional features we already know. Universities would benefit from including features such as:

- Taking into account the transfer of files with a very large size, regardless of their type;

- The possibility for a teacher or a student to share his screen with the other actors during a distance course ;

- The ability to remotely control student terminals ;

- The possibility of sharing software with students.

Distance learning also faces serious problems when it comes to assessing students' knowledge of subjects such as modern languages. It is therefore important to think about solving this problem too.

To achieve this, we use the power of the Kamailio server, the MSRP protocol [9] and the Websocket [10]. The purpose of the work is to contribute to the improvement of e-learning platforms and thus to promote better learning for students.

The organization of the article is as follows :

Section II is dedicated to the presentation of the main protocols used namely: MSRP, MSRP-RELAY, Websocket. Section III will describe how our solution works. We thus find the proposed architecture, the presentation of Kamailio and Moodle. In section IV, the results obtained from the tests will be presented. Finally, Section V concluded and gives an opening on future work.

2. Definition of MSRP, MSRP-Relay and Websocket

2.1. MSRP

The MSRP protocol is used for the exchange of instant messages and any other binary content other than ASCII (MIME)[11]. It is connection-oriented and is particularly useful for exchanging very large files or remote desktop sharing. Since MSRP[12] can not perform a session setup on its own, it is usually paired with the SIP protocol[13]. The SIP protocol is responsible for establishing and releasing the multimedia session.

MSRP defines two request types, or methods. SEND requests are used to deliver a complete message or a chunk (a portion of a complete message), while REPORT requests report on the status of a previously sent message, or a range of bytes inside a message. [8]

2.2. MSRP Relay

The MSRP protocol has NAT traversal problems. To solve this problem, the MRSP Relay is used in many implementations. [14] This is the case of our proposed solution with Kamailio. The operating principle is very simple. Customers who want to send a message can go through a relay or not. When a relay is used, it is the latter who sends the message to the customer or another relay (Figure 2).

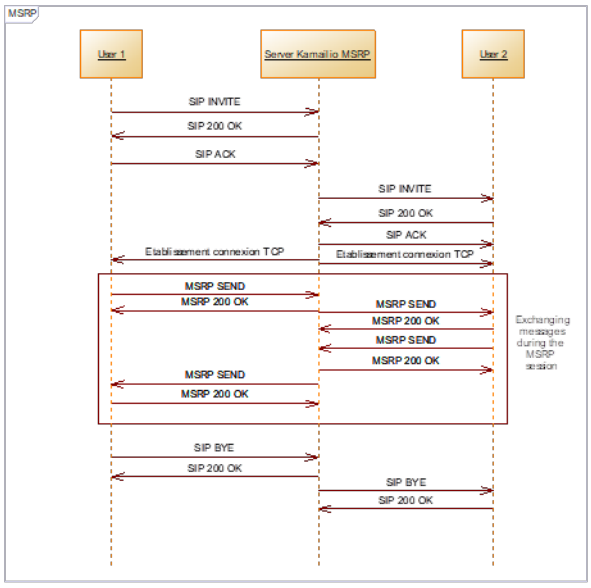


Figure 1: messaging session with MSRP

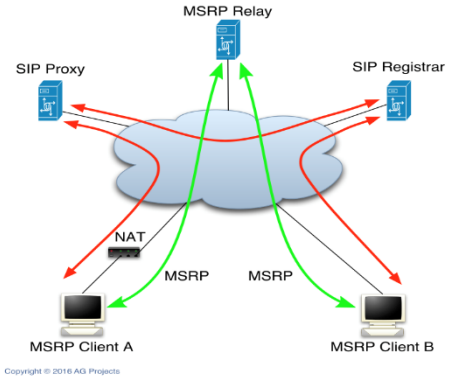


Figure 2: session message with MSRP Relay

2.3. Websocket

The WebSocket protocol enables two-way real-time communication between clients and servers. With this API you can send messages to a server and receive its responses in an events based mode without having to go to the server to get an answer [15]. The protocol has two parts : a handshake and the data transfer. Once the client and server have both sent their handshakes, and if the handshake was successful, then the data transfer part starts. This is a two-way communication channel where each side can, independently from the other, send data at will. The client can request that the server use a specific subprotocol (such as the WebSocket MSRP subprotocol) by including the "Sec-WebSocket-Protocol" field in its handshake [16].

2.4. Websocket protocol as a transport for MSRP

As the name suggests, websocket msrp subprotocol is a subset of the WebSocket protocol. He specializes in the transport of MSRP messages. These exchanges take place between a Websocket client and a node that will be called MSRP relay. It is not superfluous to wonder about the information contained in these messages. These are sometimes MSRP requests, sometimes answers. All this is

done using a Websocket connection [17] In the websocket msrp subprotocol, we find the terms clients and servers. The MSRP WebSocket Client is the entity with the ability to open multiple outbound connections to other MSRP-Relay. These other MSRP-Relay are called MSRP Websocket server. So, an MSRP WebSocket Server is simply an MSRP-Relay responsible for listening to incoming connections. These connections come from Websockets clients. Communication is obviously done with the websocket msrp subprotocol. It should first be that certain rules are respected. During the handshake procedure, both entities agree on the use of the websocket msrp subprotocol for communication. To do this, the MSRP Websocket has the obligation to put in a header called sec-websocket, the keyword 'msrp' in his handshake. Naturally, the server also in responding does the same operation in its header.

3. Architecture of our solution

3.1. Solution Architecture

The solution we proposed works as shown in Figure 3 below. The basis of architecture is the well-known moodle platform. It is to this platform that we added new features via a plugin that has been integrated. This plugin uses a free IP telephony server called Kamailio [17] whose IP address will be set in the configuration file of the module and materialized on the same figure by:

- Presentation Lib: this is the set of libraries of the plugin interacting with the MSRP API [18].
- UI presentation: represents the views of our plugin

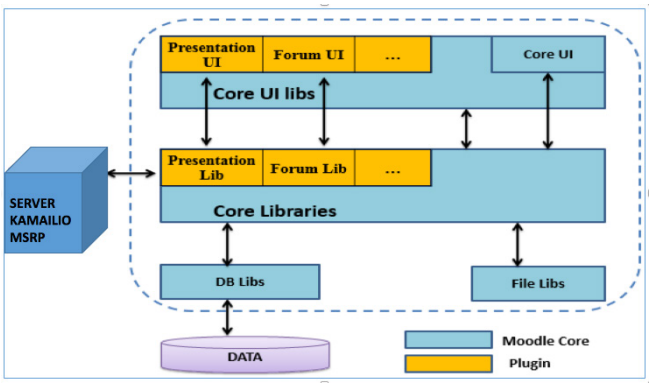


Figure 3: Architecture de la solution

3.2. How the solution works

The Websocket, MSRP and MSRP Relay protocols are the main ones involved in our platform. The solution is based on the interconnection between the kamailio server and the Moodle platform in order to push the limits of Moodlle, especially in terms of heavy file sharing or remote learning of a student's machine in order to show him good practices. The MSRP relay is mainly used to solve NAT traversal problems. Thanks to the loading and activation of the modules, routing rules have been defined especially in case of MSRP transaction. Figure 4 describes message exchanges between two users during an MSRP session. During this session, features such as screen sharing or sending large files can be used by both students and the teacher. Concretely, the user 1 sends an instant message to the user 2. The WebSocket Kamailio MSRP server receives the message and sends it to the MSRP relay which finally conveys the message to the user 2.

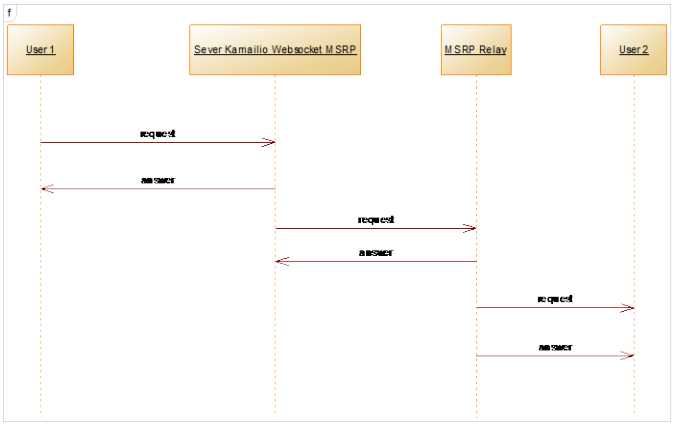


Figure 4: typical MSRP session 1

```
#!ifdef WITH_MSRP
loadmodule "msrp.so"
loadmodule "htable.so"
loadmodule "cfgutils.so"
#!endif
#!ifdef WITH_WEBSOCKETS
loadmodule "xhttp.so"
loadmodule "websocket.so"
loadmodule "nathelper.so"
```

Figure 5: configuration Server loading modules

```
event_route[msrp:frame:in] {
    msrp_reply_flags("1");

    if (($Rp == MY_WS_PORT || $Rp == MY_WSS_PORT)
            && !(proto == WS || proto == WSS || $Rp == MY_MSRP_PORT)) {
        xlog("L_WARN", "MSRP request received on $Rp\n");
        msrp_reply("403", "Action-not-allowed");
        exit;
    }

    if (msrp_is_reply()) {
        msrp_relay()
    } else if($msrp(method)=="AUTH") {
        if($msrp(nexthops)>0) {
            msrp_relay();
            exit;
        }

        if (!www_authenticate("MY_DOMAIN", "subscriber",
                                "$msrp(method)")) {
            if (auth_get_www_authenticate("MY_DOMAIN", "1",
                                "$var(wauth)")) {
                msrp_reply("401", "Unauthorized",
                                "$var(wauth)");
            } else {
                msrp_reply("500", "Server Error");
            }
            exit;
        }

        if ($hdr(Expires) != $null) {
            $var(expires) = (int) $hdr(Expires);
```

Figure 6 : MSRP transaction routing configuration

3.3. Description of Moodle

Moodle is an open source online course management system used by many organizations, universities and even businesses to offer online learning and the online component of blended learning. [19]. It is one of the most used open source course management systems in the world. The acronym MOODLE stands for "Modular Object Oriented Dynamic Learning Environment". It was in 1990 that Marin Dougiamas started designing the platform in Australia. It is an online platform under the GNU Public License (GPL). If the solution is adopted by several universities, it is mainly because it offers a set of integrated tools initially designed

on the basis of the pedagogical model called socioconstructivism [21]. Its user interface is user friendly and the solution works equally well on desktops and mobile devices. From the point of view of administration, Moodle is very ibleexible. For example, many options are available for user authentication and course registration because of the many plug-ins available. These provide organizations with many approaches to managing how they integrate Moodle into their other systems. Figure 2 below shows the architecture of Moodle.

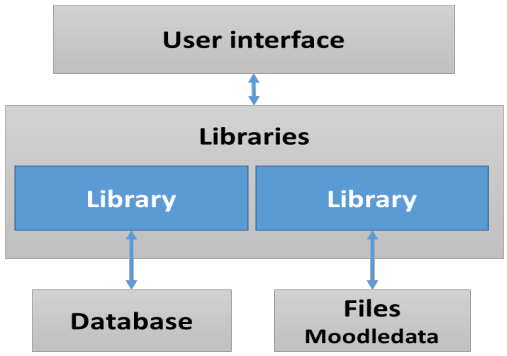


Figure 7 : Architecture de Moodle

Table 1: 1 moodle basic files

Files or Folders	Description of files or folders
version.php	It contains the following information: The declared version of the module; the minimum required version of Moodle for the plugin to work and cron settings (regular task).
index.html	Allows you to reach the list of instances of pedagogical objects in a class.
mod.html	Allows Moodle to reach the object editing screen.
icon.gif	the icon that appears in the class to symbolize the activity.
lib.php	Library functions where Moodle will look for various callbacks that provide the "standard" module information.
view.php	Allows you to make all the screens of the module.
Backup	Folder containing backup and restore information.

3.4. Moodle plugin development

A plugin or plug-in is a package that complements a host software to bring new features.

It can be positioned in a pedagogical sequence of a course. The following table shows some basic files of a moodle plugin.

3.5. Description of Kamailio

Kamailio is a very powerful open source SIP server, capable of handling thousands of calls per second without degradation of the quality of service. It can be used to create real-time VoIP communication platforms. It offers multiple features, including support for WebSocket for WebRTC, IPv4 and IPv6, TCP, UDP, TLS for secure communication, IMS extensions, and even an MSRP relay module discussed in this article. The MSRP relay module used in this work reuses the Kamailio core module to provide MSRP routing capabilities. A not insignificant advantage of this module is that it can access all other extensions of the Kamailio server. These are authentication, accounting, security and protection against denial of service attacks, etc. Kamailio has the ability to handle both SIP traffic and SIP traffic received on the same port. This is how Kamailio is used as a standalone MSRP relay where you can have both SIP and MSRP processing instances. You can also configure the module so that some ports will be for SIP only and others for MSRP.

4. Results

To prove the relevance of our solution, we deployed and tested it with 22 students and obtained the following results. Especially in this extended version, we were able to give access to a paid software to create architectures with several licenses. Thus, thanks to a single license, students can connect remotely and use the software to do their practical work, practice without the help of the teacher, etc.

4.1. Screen sharing

Figure 8 shows the invitation for screen sharing with the three options accept, reject, and occupy.

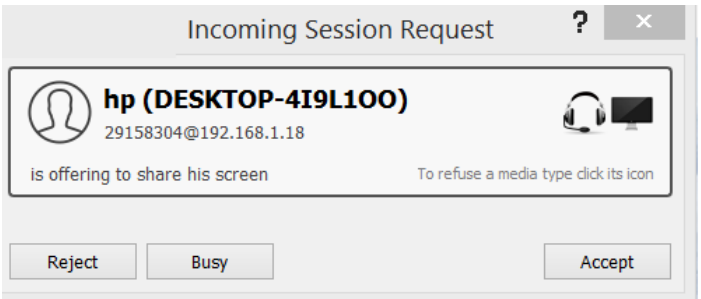


Figure 8: incoming session request

Figure 9 shows the sharing and remote control of a computer between trainer and learners. The teacher asks to take control of the learner's machine in order to show him how to handle.



Figure 9: Sharing and remote control of a computer

Figure 10 shows an illustration of the sharing and remote control of a workstation between actors in a digital university. This feature makes possible to pass audio video streams and even chat.

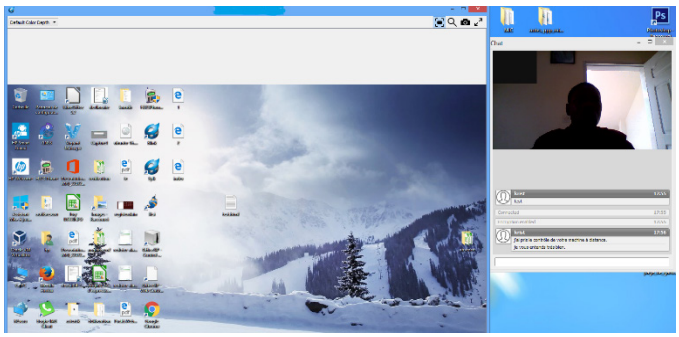


Figure 10 : remote sharing and control integrating audio and video

The importance of practical work in teaching in STEM is no longer demonstrated. An environment rich in practical work ensures better education for students. Relatively expensive software required for student training may not be accessible to them remotely. Thanks to our platform, students can access virtual machines containing these programs. Figure 11 shows how students can access these virtual machines through our platform. Thus, the student will be able to practice freely by choosing the virtual machine containing the software which he needs.

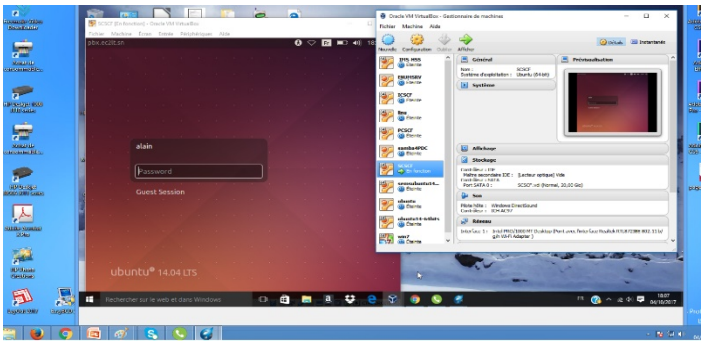


Figure 11: Remote access to a virtualization server for hands-on labs

Figure 12 shows a specific work environment for students learning computer programming. Thanks to a collaborative development environment from a web browser, students can work on the same project together. It may be to develop a project together using the eclipse development environment for example.

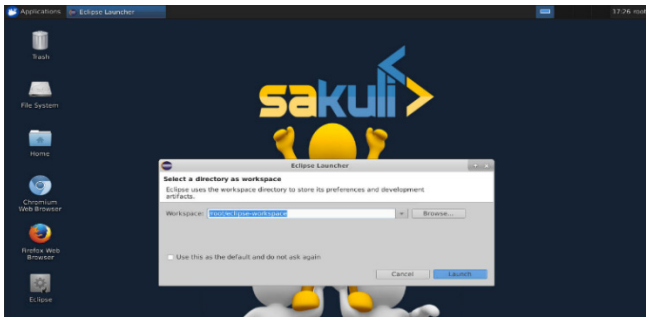


Figure 12: sharing development environment

In a classical classroom training, the teacher sometimes assists a student in difficulty. He can therefore show him step by step how to use this or that feature. This practice should also be possible in distance learning. This is what we demonstrate with our platform (Figure 13). A student's machine was remote-controlled by the teacher. The latter shows his student some features specific to a scientific computing software.

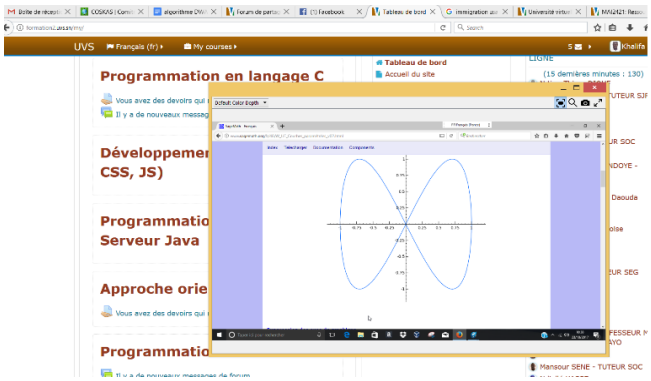


Figure 13: remote control of a learner's machine

Some programs are essential for learners as part of their training (development of architectures, mathematical simulation software, etc.). Sometimes, these tools are not free and not all students can acquire the license. To help the students, we provide them with a virtual machine on which we install the tools and buy the license. Thus, with a single license, during a session all students in the class will be able to use the software.

Figure 14 shows the E-DRAW software for example, used by computer science students to create diagrams, architectures, etc. Each student can therefore connect remotely, use the software to draw his architectures and save his work.

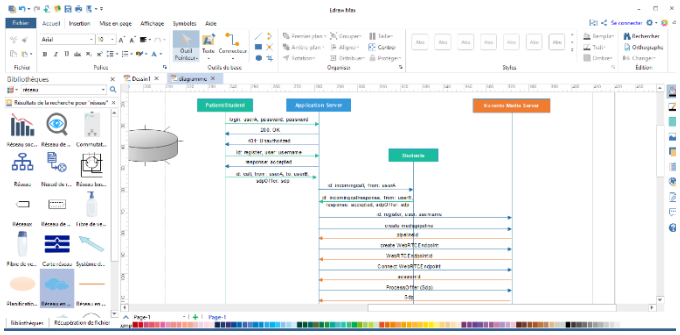


Figure 14: remote use of edraw

We also note that students at the beginning of their studies have difficulty installing certain software because of the limited resources of their computers. So we made SageMath software available to students to allow them to do their simulations remotely without worrying about the performance of their computer.

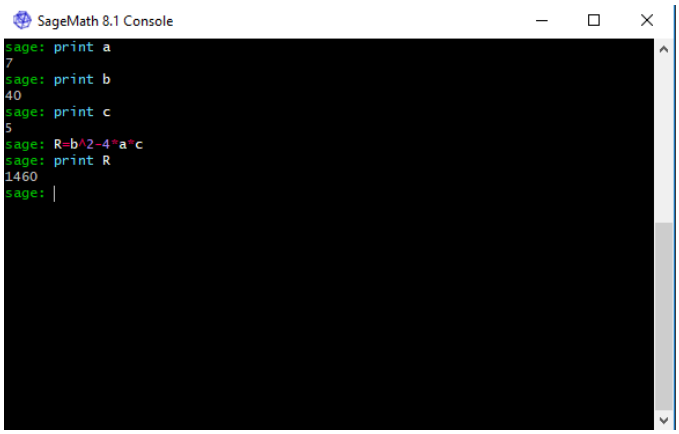


Figure 15: remote use of SageMath

4.2. Transfer of large files

This is about sending very heavy files simply by using drag and drop. It should be specified that the one to whom the file is intended has the ability to accept or reject the heavy file sent. If it chooses to accept the file, the download will start as you can see in Figure 16.

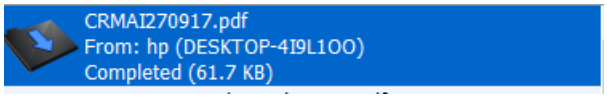


Figure 16 file receive 1

5. Conclusion and perspective

In this work, we proposed the integration of innovative tools in distance learning devices, based on the protocols WebSocket, MSRP, and the Kamailio media server, to allow the different actors of virtual universities to improve the quality of practical work in online labs. Students from the Virtual University of Senegal were able to use this solution effectively to collaborate in real time on practical work as if they were physically on the Moodle platform. It should be noted that the students of the Virtual University of Senegal are already used to the use of Moodle.It also allows the various actors of the training device to exchange large files thanks to the flexibility of the MSRP protocolThe platform has also been experimented with math students for online tutorials. The result was very satisfying. Two major features in this solution have been very useful for language teachers. This is audio communication and large file transfer in real time. It was helpful for them to truly appreciate the reading and pronunciation skills of their students.

The websocket msrp subprotocol also had a real impact in the implementation of the solution. Indeed, it made it easy to integrate our solution as a Moodle plugin. There was no question of changing the habits of students who are already accustomed to the Moodle GUI.

In our experiments, we used twenty-four students, which allowed us to easily manage the quality of service to users. We plan to use the Computing Cloud and a load-balancing device to manage the scalability, allowing the MOOCs to use our solution.

Acknowledgment

The authors wish to express their sincere thanks to members of the Computer, Networks and Telecoms Laboratory of UCAD ESP and all those who contributed to the success of this work.

References

[1] K. Sylla, M. Seck, S. Ouya and G. Mendy, "Impact of MSRP protocol integration in e-learning platforms of universities," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon-si Gangwon-do, Korea (South), 2018, pp. 1-1.

[2] A. Sharma, "A proposed e-learning system facilitating recommendation using content tagging and student learning styles," 2017 5th National Conference on E-Learning & E Learning Technologies (ELELTECH), Hyderabad, India, 2017, pp. 1-6.

[3] S.B. Nite ; M. Margaret ; R.M. Capraro ; J. Morgan; C.A. Peterson, "Science, technology, engineering and mathematics (STEM) education: A longitudinal examination of secondary school intervention,"Frontiers in Education Conference (FIE), 2014 IEEE , vol., no., pp.1,7, 22-25 Oct. 2014.

[4] J. P. López, A. Cerezo, J. M. Menéndez and J. P. Ballesteros, "Usage of mobile devices as collaborative tools for education and preparation of official exams," 2015 International Symposium on Consumer Electronics (ISCE), Madrid, 2015, pp. 1-2.

[5] C.C. Aydin ; G. Tirkes, "Open source learning management systems in elearning and Moodle," Education Engineering (EDUCON), 2010 IEEE , vol., no., pp.593,600, 14-16 April 2010

[6] S. Ouya, A. B. Mbacke, G. Mendy, P. W. Diouf and K. Sy, "Social network integration to e-learning environment," 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), Marrakech, 2015, pp. 1-4.

[7] S. Ouya, K. Sylla, P. M. D. Faye, M. Y. Sow and C. Lishou, "Impact of integrating WebRTC in universities' e-learning platforms," 2015 5th World Congress on Information and Communication Technologies (WICT), Marrakech, 2015, pp. 13-17.

[8] M. I. Santally, R. K. Sungkur, N. Sooltangos, P. Mounier and R. Mira-Bahadoor, "Flash to HTML5 adaptations: A case-study for UpToTen.com," 2016 International Conference on Advances in Computing and Communication Engineering (ICACCE), Durban, South Africa, 2016, pp. 375- 380.

[9] B. Campbell ; R. Mahy and C. Jennings, "The Message Session Relay Protocol (MSRP)", RFC 4975, DOI 10.17487/RFC4975, September 2007, <https://www.rfc-editor.org/info/rfc4975>.

[10] V. N. Gridin, G. D. Dmitrevich, V. I. Anisimov and S. A. Vasiliev, "Methods of constructing service-oriented computer — Aided circuit design systems based on WebSocket Protocol and Oracle database," 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, 2017, pp. 409-412.

[11] Ned. FREED, RFC2045 multipurpose Internet mail extensions (MIME) part one: format of Internet message bodies. http://ds. internic. net/rfc/rfc2045. txt, 1996.

[12] C. Jennings ; R. Mahy and R. Adam. "Relay Extensions for the Message Sessions Relay Protocol (MSRP)." (2007).

[13] H. Yang, H. Lin, J. Li and W. Lei, "The Design and Implementation of an Enhanced SIP ALG to Support MSRP Sessions," 2009 Ninth International Conference on Hybrid Intelligent Systems, Shenyang, 2009, pp. 289-292.

[14] C. Jennings, R. Mahy and A. Roach. "Relay Extensions for the Message Sessions Relay Protocol (MSRP)." (2007).

[15] I. Fette and A. Melnikov, "The WebSocket Protocol", RFC 6455, DOI 10.17487/RFC6455, December 2011, <https://www.rfc-editor.org/info/rfc6455>.

[16] P. Dunkley ; G. Llewellyn ; V. Pascual ; G. Salgueiro and R. Ravindranath, "The WebSocket Protocol as a Transport for the Message Session Relay Protocol (MSRP)", RFC 7977, DOI 10.17487/RFC7977, September 2016, <https://www.rfc-editor.org/info/rfc7977>.

[17] VOZNAK, Miroslav and L. MACURA, Kamailio syntax generator and configuration file parser. In: Proceedings of the 15th WSEAS International Conference on Computers, WSEAS, Wisconsin, Estados Unidos. 2011. p. 308-312.

[18] H. M. Rissanen, T. Mecklin and M. Opsenica, "Design and Implementation of a RESTful IMS API," 2010 6th International Conference on Wireless and Mobile Communications, Valencia, 2010, pp. 86-91.

[19] G. Sykamiotis, A. Charitopoulos, M. Rangoussi and D. Koulouriotis, "Extraction and presentation of access and usage data from an e-learning platform (moodle): Design and development of a software application," 2017 IEEE Global Engineering Education Conference (EDUCON), Athens, 2017, pp. 1020-1026.

[20] M. Ivanova, G. Grosseck and C. Holotescu, "Researching data privacy models in eLearning," 2015 International Conference on Information Technology Based Higher Education and Training (ITHET), Lisbon, 2015, pp. 1-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 327-340 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Minimizing Time Delay of Information Routed Across Dynamic Temporal Sensor Networks

Michael J. Hirsch[*,1], Azar Sadeghnejad[2]

[1]*ISEA TEK, LLC, 32751, U.S.A.*

[2]*Radiation-Oncology Department, University of Texas - Southwestern Medical Center, 75390, U.S.A.*

ARTICLEINFO

Article history:
Received: 13 July, 2018
Accepted: 10 August, 2018
Online: 20 August, 2018

Keywords:
Information Routing
Sensor Network
Communication Topology

ABSTRACT

In this research we address the problem of routing information across dynamic temporal sensor networks. The goal is to determine which information, generated by sensors on resources at various times, is able to be routed to other resources, consumer resources, within the given information time window, while being constrained by temporally dynamic bandwidth limitations across the sensor network, and storage limitations on the resources. A mathematical model of the problem is derived, and used to find solutions to the problem. In addition, a heuristic is developed to efficiently find good quality solutions. Monte-Carlo simulations are performed comparing solutions found by commercial software with the heuristic.

1 Introduction

This paper is an extension of work originally presented at the Military Communications Conference [1]. Wireless sensor networks are a class of networks in which some or all of the sensors or resources collect, analyze, and communicate data acquired from their environment (external or internal) to other nodes in the network [2]. These other nodes could be resources, sensors, or fixed or mobile command posts. Mobile ad-hoc networks (MANETs) are a subset of wireless sensor networks that have an absence of a fixed infrastructure, and exhibit a dynamic communication topology [3]. Resources within transmission range of others communicate directly (depending on bandwidth limitations) and cooperation among intermediate resources is required for resources to communicate with others outside of their direct communication range. More and more networks, in both the military and commercial domain, are relying on wireless sensor networks and MANETs [4, 5, 6].

The ability for data to be collected by manned and unmanned vehicles engaged in military, government, and commercial missions is increasing at an exponential rate [7]. Examples include high-resolution video, telephone intercepts, traffic congestion reports, internal vehicle system states, among others. In commercial telecommunication networks, there are instances where the capacity across the network is in the terabytes per second range [8]. Networks of this type have the ability to transmit all data that is collected without restrictions or delays. However, in times of crisis and disaster, commercial and civil communication networks can be severely degraded [9, 10]. And in military missions, bandwidth is oftentimes severely limited, which results in a reduction in transmission and sharing of information that can be mission-critical. And this will continue to be the case due to the rise of operations performed in anti-access/area denial (A2AD) environments, against near-peer adversaries. In the past, this meant that information collected by vehicles would not be processed, analyzed, and understood until after the vehicle had returned to base and 'uploaded' all of the collections [7, 11].

In performing more of the processing and analysis on-board the vehicles, there has the potential for less bandwidth requirements for vehicles to share information important for mission success. In effect, this can be seen as *pushing the smarts out to the tactical edge*. Each generation of manned and unmanned vehicles is having more and more processing capabilities on-board [11]. Examples include internally monitoring on-board systems, fusing information collected from various on-board sensors, picking out a person of interest from a full motion video, and deciding how much information should be sent over the sensor network and to whom should be the recipient of such information [12]. While some of the collected information can

[*]Michael J. Hirsch, ISEA TEK, 620 N. Wymore Rd., Ste. 260, Maitland, FL 32751 U.S.A., mhirsch@iseatek.com

https://dx.doi.org/10.25046/aj030433

just be generated and stored on a vehicle, and offloaded upon mission completion, other generated information is vital to mission success, and must be transmitted to control station(s) or other resources during mission execution.

To address this, we consider the following problem. There are a set of resources (moving or fixed) that have a set of Information Generation Requests (IGRs) to satisfy over a given time horizon. If a resource executes a task to perform the IGR, then an associated piece of information is generated. We will use the term IGR to denote the information that gets generated. Each IGR can come from one of a fixed set of categories (for an example of information categories, see Table 1). In addition, each IGR has associated with it the following characteristics: an expected time of generation, the set of consumer resources (those resources that would find this information important for their mission), the time by which the information needs to be delivered to the to be of use, the expected original size of the information, and the priority of the information (i.e., a measure of its importance in the mission). We note that the priority of an IGR could be determined *a priori* (e.g., information collected about one kinematic location is deemed very important and therefore is given a high priority) or dynamically on-board the resource at the time of collection, if the resource has the appropriate analytics (e.g., image is given low priority because it does not contain any red cars). By a 'consumer resource' for a particular IGR, we mean that the resource actually performs some processing with that information (e.g., fuses that information with other information [13, 14], makes command and control decisions, etc.).

There is a dynamic sensor network topology in place over the time horizon. This dynamic network has finite, but varying bandwidth limitations between any two resources at any given time, and in many cases there might not be a direct connection between two resources. Each IGR can be sent over the network with different potential granularities (e.g., 100% of original size, 70% of original size, etc.), but there is a minimum size percentage with which the information remains useful to the consumer of the information. As an IGR gets routed from the generating resource to the consumer resource, all resources along this route will need to store this IGR locally for some amount of time. The generating resource will need to store an IGR from the time of generation until the IGR is completely sent to the next resource in the path to the consumer resource(s). Non-consumer resources along this path will need to store the IGR from the time the resource begins receiving the IGR until the time the resource has completely transmitted the IGR to the next resource along the path. And the consumer resource(s) will need to store the IGR form the time they begin receiving it until at least they have finished processing the IGR.

Each resource has a finite storage capacity, and resources can triage information they have stored (e.g., once a piece of information is generated and completely sent to the next resource along the path to the consumer, the generator resource can delete the information from storage). The objective is then to determine which IGRs to route over the network, how to route those IGRs so that they have a minimal time delay (difference between time of consumption and time of generation), from generating resource to consumer resource over the time horizon, while also attempting to maximize the size of granularity of each IGR that is routed. In turn, this produces the higher-level end product of determining which IGRs should be accomplished for mission objectives, given the network topology characteristics.

There has been some prior research into routing schemes for collected data across a dynamic network. In [17], the author considered the problem of sinkholes in a network. A sinkhole node attempts to deceive all the nodes in the network to route network traffic to the sinkhole, by broadcasting false routing information across the network. An approach based upon secondary caching was developed to prevent a sinkhole attack in dynamic source routing sensor networks. In [18], the author consider cellular data networks and the significant increase in data being transmitted over these networks. Their research looked at data traffic management techniques, whereby users and automated approaches 'flag' messages that are of low priority, and an higher-level agent-based scheduling system had the ability to delay scheduling of low priority messages during peak load times. In [19], the authors considered the problem of extending lifetimes of wireless sensor networks. In their view, the sending of redundant data across sensor networks reduced the lifetime of these networks. They looked at providing certain nodes in the network to function as 'data aggregation' nodes, receiving collected data from other nodes in the network, fusing or aggregating the data together to reduce redundant information, and then passing this reduced set of data onwards to a base station. They employed a grid-based routing and aggregator selection scheme to find the minimum number of aggregation points while routing data to the base station ensuring that the network lifetime is maximized.

In addition, there has been some prior research addressing certain aspects of our problem of interest, but none that has addressed the problem completely. In [20], the author considered the problem of routing information across a network, from a workflow perspective, but the network was assumed to be complete and static, and bandwidth was not addressed. In [21], the researchers extended this work to include bandwidth limitations, but still under the assumption on a complete and static network. In [22], the researchers presented a mixed-integer linear programming formulation to model the trajectory of a set of unmanned aerial vehicles, along with routing of information collected from these vehicles back to a base station. They did not consider bandwidth fluctuations in the sensor network, and the resources were not the recipient of any information, just the base station. In addition, all of their information was considered to have the same priority. In [23], the author considered underwater

Table 1: Examples of Information Categories.

Category	Example
Internal System State	Measurements on the engine pressure ratio [15]
External Collection	Photographs of an urban traffic intersection
Situation Assessments	High likelihood that red truck contains person of interest
Command and Control (C2)	Modifying flight plan to conserve battery life [16]

acoustic networks and the long propagation delays and low bandwidths inherent in such networks. A set of nodes was using the same bandwidth channel, and because their problem had bursty data, the solution approach looked at submitting their requests randomly, as demand dictated. A medium access control protocol was in place to minimize the number of collisions among the information flowing across the network. In [24], the researchers investigated ways to enhance the operations of a power grid. One of the ways, taking advantage of the high sampling rates of the measurement data, required a high bandwidth, networked communication system. Their results derived a method to simulate, design, and test the adequacy of a communication system for a particular grid layout.

The rest of this paper is organized as follows. In Section 2, we introduce the parameters and decision variables for our model, and rigorously derive the mathematical program representing our problem. Section 3 develops a heuristic to efficiently find good quality solutions. Section 4 analyzes the results of computational experiments performed on various scenarios, comparing commercial software solving a linearized version of the mathematical program and the solutions found by the heuristic. Section 5 provides some concluding remarks and future research directions.

2 Mathematical Formulation

In this section, we rigorously derive mathematical equations to model the problem of interest. This results in a mixed-integer nonlinear program (MINLP). We note that the time dimension in the model is discretized, over the planning horizon of $\mathcal{K}$ time-steps.

2.1 Parameters

This section lists the parameters for the mathematical programming problem.

$\mathcal{I}$ is the set of resources, $i, i_1, i_2, \in \mathcal{I}$;

$\mathcal{J}$ is the number of Information Generation Requests (IGR)s, $j \in \{1, \ldots, \mathcal{J}\}$;

$\mathcal{K}$ is the number of time-steps in the planning horizon, $k, \hat{k} \in \{1, \ldots, \mathcal{K}\}$;

$\mathcal{P}^{min}$ is the minimum percentage an IGR can be decreased in size and still be considered useful to consumer resource(s);

$\tau_{i,j}$ is the number of time-steps after a consumer receives an IGR before the consumer can delete the IGR, i.e., the number of time-steps it takes resource i (the consumer resource) to 'process' IGR j;

ρ_j is the expected priority of IGR j;

S_j^o is the expected original size of IGR j;

g_j is the expected time of generation of IGR j;

d_j is the time by which IGR j is due to the consumer resource(s) of the IGR;

$$G_{i,j} = \begin{cases} 1 & \text{if resource } i \text{ is tasked with generating IGR } j \\ 0 & \text{o.w.} \end{cases};$$

$$C_{i,j} = \begin{cases} 1 & \text{if resource } i \text{ is a consumer of IGR } j \\ 0 & \text{o.w.} \end{cases};$$

L_i is the storage capacity of resource i;

$b_{i_1,i_2,k}$ is the expected bandwidth capacity from resource i_1 to resource i_2 at time-step k;

$\bar{\rho}$ is the largest expected priority, i.e., $\bar{\rho} = \max_j \left[\rho_j\right]$;

ω_d and ω_s are the weighting coefficients for the priority and size components of the objective function;

$\mathcal{M}$ is a large enough constant, used in Constraints (7).

2.2 Variables

This section lists the decision variables for the mathematical programming problem.

$$x_j = \begin{cases} 1 & \text{if IGR } j \text{ is chosen to be routed} \\ & \text{from the generating resource to} \\ & \text{the consumer resource(s)} \\ 0 & \text{o.w.} \end{cases}, \quad \forall j;$$

s_j = the size of IGR j sent to the consumer resource(s), $\forall j$;

$a_{i_1,i_2,j,k}$ = the amount (size) of IGR j sent from resource i_1 to resource i_2 during time-step k, $\forall i_1, i_2 (\neq i_1), j, k$;

$$\alpha_{i_1,i_2,j,k} = \begin{cases} 1 & \text{if resource } i_1 \text{ sends some of IGR } j \\ & \text{to resource } i_2 \text{ during time-step } k \\ 0 & \text{o.w.} \end{cases},$$

$\forall i_1, i_2 (\neq i_1), j, k$;

$m_{i,j,k}$ = the amount of IGR j being stored on resource i at time-step k, $\forall i,j,k$. **N.B.**: For consistency of the constraints, $m_{i,j,0} = 0 \ \forall i,j$;

$$\phi_{i_1,i_2,j} = \begin{cases} 1 & \text{if resource } i_1 \text{ is assigned to send IGR } j \text{ to resource } i_2 \\ 0 & \text{o.w.} \end{cases},$$

$$\forall i_1, i_2 (\neq i_1), j;$$

$\beta_{i,j}$ = the amount of time resource i takes in sending IGR j to other resources, i.e., the time between resource i completely receiving or generating IGR j to the time where resource i no longer needs IGR j, $\forall i,j$;

$$t_{i,j,k} = \begin{cases} 1 & \text{if resource } i \text{ cumulatively receives (or generates) } s_j \text{ size units of IGR } j \text{ at time-step } k \\ 0 & \text{o.w.} \end{cases},$$

$$\forall i,j,k;$$

$T_{i,j}$ = the time-step at which resource i completely receives (or generates) IGR j, $\forall i,j$;

$$d_{i,j,k} = \begin{cases} 1 & \text{if resource } i \text{ deletes IGR } j \text{ from storage at time-step } k \\ 0 & \text{o.w.} \end{cases},$$

$$\forall i,j,k;$$

$D_{i,j}$ = the time-step at which resource i deletes IGR j from its storage, $\forall i,j$;

$Q_{i,j}$ = the time at which resource i completely consumes (or transmits) IGR j, $\forall i,j$;

$$e_{i,j,k} = \begin{cases} 1 & \text{if resource } i \text{ receives (or generates) IGR } j \text{ by time-step } k \\ 0 & \text{o.w.} \end{cases},$$

$$\forall i,j,k.$$

N.B.: For consistency of the constraints, it is necessary to set $e_{i,j,0} = 0 \ \forall i,j$. Also, note that $e_{i,j,k} = 1$ implies that $e_{i,j,\ell} = 1$ for each $\ell > k$ and $e_{i,j,k} = 0$ implies that $e_{i,j,\ell} = 0$ for each $\ell < k$.

2.3 Nonlinear Mathematical Formulation

This section presents the mathematical model of the problem described in Section 1, resulting in a mixed-integer nonlinear program (MINLP).

$$\max \ \omega_d \cdot \left(\sum_{j=1}^{\mathcal{J}} \sum_{\substack{i \in \mathcal{I}: \\ C_{i,j}=1}} \frac{\left(\mathcal{K} + g_j\right) \cdot x_j - T_{i,j}}{\mathcal{K}} \right) + \omega_s \cdot \left(\sum_{j=1}^{\mathcal{J}} \frac{s_j}{S_j^o} \right) \quad (1)$$

s.t.

$$s_j \leq S_j^o \cdot x_j \ \ \forall j \quad (2)$$

$$s_j \geq \mathcal{P}^{min} \cdot S_j^o \cdot x_j \ \ \forall j \quad (3)$$

$$\sum_j m_{i,j,k} \leq L_i \ \ \forall i,k \quad (4)$$

$$m_{i,j,k} = m_{i,j,k-1} + \sum_{\substack{i_1 \in \mathcal{I} \\ i_1 \neq i}} a_{i_1,i,j,k} - d_{i,j,k} \cdot s_j + \delta\left[g_j - k\right] \cdot G_{i,j} \cdot s_j \ \ \forall i,j,k \quad (5)$$

$$\sum_{k=1}^{\mathcal{K}} \alpha_{i_1,i_2,j,k} \leq \mathcal{K} \cdot x_j \ \ \forall i_1, i_2, j \quad (6)$$

$$\mathcal{M} \cdot \alpha_{i_1,i_2,j,k} \geq a_{i_1,i_2,j,k} \ \ \forall i_1, i_2, j, k \quad (7)$$

$$\sum_j a_{i_1,i_2,j,k} \leq b_{i_1,i_2,k} \ \ \forall i_1, i_2, k \quad (8)$$

$$\sum_{k=1}^{g_j} a_{i_1,i_2,j,k} = 0 \ \ \forall i_1, i_2, j \quad (9)$$

$$\sum_{k=1}^{\mathcal{K}} a_{i_1,i_2,j,k} \leq s_j \ \ \forall i_1, i_2, j \quad (10)$$

$$e_{i,j,k} > \sum_{\substack{i_1 \in \mathcal{I} \\ i_1 \neq i}} \sum_{\hat{k}=1}^{k} a_{i_1,i,j,\hat{k}} - s_j - \left(1 - x_j\right) \ \ \forall k, \ \forall i,j \text{ s.t. } G_{i,j} = 0 \quad (11)$$

$$e_{i,j,k} \cdot s_j \leq \sum_{\substack{i_1 \in \mathcal{I} \\ i_1 \neq i}} \sum_{\hat{k}=1}^{k} a_{i_1,i,j,\hat{k}} \ \ \forall k, \ \forall i,j \ s.t. \ G_{i,j} = 0 \quad (12)$$

$$t_{i,j,k} \geq e_{i,j,k} - e_{i,j,k-1} \ \ \forall i,j,k \quad (13)$$

$$\sum_{k=1}^{\mathcal{K}} t_{i,j,k} \leq x_j \ \ \forall i,j \quad (14)$$

$$t_{i,j,g_j} = x_j \ \ \forall i,j \ s.t. \ G_{i,j} = 1 \quad (15)$$

$$T_{i,j} = \sum_{k=1}^{\mathcal{K}} k \cdot t_{i,j,k} \ \ \forall i,j \quad (16)$$

$$T_{i,j} \leq d_j \cdot x_j \ \ \forall i,j \quad (17)$$

$$Q_{i,j} = T_{i,j} + \tau_{i,j} \cdot x_j \ \ \forall i,j \ s.t. \ C_{i,j} = 1 \quad (18)$$

$$Q_{i,j} = T_{i,j} + \beta_{i,j} \ \ \forall i,j \ s.t. \ C_{i,j} = 0 \quad (19)$$

$$\beta_{i_1,j} \geq k \cdot \alpha_{i_1,i_2,j,k} - T_{i_1,j} \ \ \forall i_1, i_2, j, k \quad (20)$$

$$D_{i,j} > Q_{i,j} - \left(1 - \sum_{k=1}^{\mathcal{K}} t_{i,j,k}\right) \ \ \forall i,j \quad (21)$$

$$D_{i,j} = \sum_{k=1}^{\mathcal{K}} k \cdot d_{i,j,k} \quad \forall i,j \tag{22}$$

$$\sum_{k=1}^{\mathcal{K}} d_{i,j,k} \leq x_j \quad \forall i,j \tag{23}$$

$$\sum_{\substack{i_1 \in \mathcal{I} \\ i_1 \neq i}} \phi_{i_1,i,j} \leq x_j \quad \forall i,j \tag{24}$$

$$\sum_{k=1}^{\mathcal{K}} \alpha_{i_1,i_2,j,k} \leq \mathcal{K} \cdot \phi_{i_1,i_2,j} \quad \forall i_1,i_2,j \tag{25}$$

$$\sum_{k=1}^{\mathcal{K}} \alpha_{i_1,i_2,j,k} \geq \phi_{i_1,i_2,j} \quad \forall i_1,i_2,j \tag{26}$$

$$\sum_{\substack{i_2 \in \mathcal{I} \\ i_2 \neq i}} \alpha_{i,i_2,j,k} \leq \sum_{\hat{k}=1}^{k-1} t_{i,j,\hat{k}} \quad \forall i,j,k \tag{27}$$

$$s_j = \sum_{k=1}^{\mathcal{K}} \sum_{\substack{i_1 \in \mathcal{I} \\ i_1 \neq c_j}} a_{i_1,c_j,j,k} \quad \forall j \tag{28}$$

$$\sum_{k=1}^{\mathcal{K}} \sum_{i_2 \in \mathcal{I}} \alpha_{c_j,i_2,j,k} = 0 \quad \forall j \tag{29}$$

$$x_j, \alpha_{i_1,i_2,j,k}, \phi_{i_1,i_2,j}, t_{i,j,k}, d_{i,j,k}, e_{i,j,k} \in \{0,1\} \quad \forall i,i_1,i_2,j,k \tag{30}$$

$$s_j, a_{i_1,i_2,j,k}, m_{i,j,k} \in \left[0, S_j^o\right] \quad \forall i,i_1,i_2,j,k \tag{31}$$

$$\beta_{i,j}, D_{i,j}, T_{i,j}, Q_{i,j} \in [0,\mathcal{K}] \cap \mathbb{Z} \quad \forall i,j,k \tag{32}$$

2.4 Interpretation of Nonlinear Mathematical Formulation

In this section, we explain the objective function and each of the constraints that are part of the mathematical formulation.

- The objective function, Equation (1), is a weighted combination of the time delay and size of the IGRs routed across the network.

  The first term in the objective function deals with the time delay of those IGRs chosen to be routed. The sum is over all IGRs j and all resources i that are consumers of IGR j (i.e., those resources such that the parameter $C_{i,j} = 1$). When an IGR j is not chosen to be routed, x_j is set to be 0 and $T_{i,j}$ also ends up as 0, so this term is 0. When an IGR is chosen to be routed, x_j is set to be 1 and $T_{i,j}$ is set to be the time at which resource i completely receives IGR j. The time IGR j is generated is given by the parameter g_j. The maximum value of this term is 1, which occurs when $T_{i,j} = g_j$ (which is an idealistic situation), while the minimum value of this term is $\frac{\mathcal{K}+g_j-d_j}{\mathcal{K}}$, which can be 0 when g_j is 0 and d_j is $\mathcal{K}$.

  The second term of the objective function deals with the size of the IGRs chosen to be routed, looking at the ratio of the actual size of the IGR that gets routed versus the original size.

  Note that that scale of these two terms are the same, i.e., they are both within $[0,1]$, and are unit-less. The first term is weighted by ω_d, while the second term is weighted by ω_s.

- Constraints (2) force s_j to be less than or equal to S_j^o if IGR j is chosen to be routed, and 0 if IGR j is not chosen to be routed;
- Constraints (3) require s_j to be greater than or equal to the minimal amount to be routed for IGR j ($\mathcal{P}^{min} \cdot S_j^o$), if IGR j is chosen to be routed;
- Constraints (4) ensure that at each time-step, the storage capacity for each resource is not exceeded;
- Constraints (5) determine the amount of IGR j stored on resource i at time k as the amount stored at time $k-1$ plus the amount sent to resource i at time-step k minus s_j if IGR j is deleted from resource i's storage at time-step k, plus s_j if IGR j is chosen to be routed by resource i and is generated at time-step k. **N.B.**: The Dirac-Delta impulse function $\delta[q]$ returns 1 when $q = 0$ and returns 0 otherwise [25];
- Constraints (6) only allow $\alpha_{i_1,i_2,j,k}$ to be greater than 0 if IGR j is chosen to be routed;
- Constraints (7) force $a_{i_1,i_2,j,k}$ to be 0 if resource i_1 is not sending IGR j to resource i_2 at time-step k;
- Constraints (8) limit the amount of information traveling from resource i_1 to i_2 at time-step k to be no more that the capacity of this edge, $b_{i_1,i_2,k}$;
- Constraints (9) enforce that no resource can send IGR j before it is generated;
- Constraints (10) enforce that resource i_1 send no more than s_j of IGR j to resource i_2;
- For those resources that do not generate IGR j and IGR j is chosen to be routed, Constraints (11) force $e_{i,j,k}$ to be 1 if all s_j of IGR j reaches resource i at or before time-step k. This constraint allows $e_{i,j,k}$ to be 0 or 1 otherwise;
- For those resources that do not generate IGR j, Constraints (12) force $e_{i,j,k}$ to be 0 while not all s_j of IGR j has reached resource i. This constraint allows $e_{i,j,k}$ to be 0 or 1 otherwise;
- Constraints (13) set $t_{i,j,k}$ to be 1 at the time-step that resource i receives all s_j of IGR j, and to be 0 for all other time-steps;
- Constraints (14) enforce that resource i can receive IGR j at most once;
- Constraints (15) set $t_{i,j,k}$ equal to 1 if IGR j is chosen to be routed, is generated by resource i, and k is the time of generation of IGR j;

- Constraints (16) set $T_{i,j}$ to be the time at which resource i receives IGR j;
- Constraints (17) enforce that any resource that receives IGR j, must do so before its due date/time, if IGR j is chosen to be routed;
- Constraints (18) define the time at which resource i has completely 'processed' IGR j, if resource i is a consumer of IGR j and IGR j is chosen to be routed;
- Constraints (19) define the time at which resource i has completely transmitted IGR j to other resources, if resource i is not a consumer of IGR j and IGR j is chosen to be routed;
- Constraints (20) determine the amount of time it takes resource i to send IGR j to other resources, once resource i receives or generates IGR j;
- Constraints (21) ensure that if IGR j is chosen to be routed, that IGR j cannot be deleted from resource i until resource i is finished with the IGR;
- Constraints (22) determine the time at which resource i deletes IGR j;
- Constraints (23) ensure that resource i can delete IGR j at most once;
- Constraints (24) enforce that at most one resource is assigned to send IGR j to resource i;
- Constraints (25) only allow resource i_1 to send IGR j to resource i_2 if it is assigned to;
- Constraints (26) force i_1 to send IGR j to resource i_2 during at least one time-step, if i_1 is assigned to send IGR j to resource i_2;
- Constraints (27) does not allow resource i to send any portion of IGR j at time-step k if resource i has not completely received (or generated) IGR j by time-step $k-1$;
- Constraints (28) ensure that the consumer resource(s) of IGR j receives all s_j units, if IGR j is chosen to be routed;
- Constraints (29) prohibit resource c_j (the consumer of IGR j) from sending IGR j to other resources;
- Constraints (30) – (32) are domain restriction constraints on the decision variables.

3 Solution Methodologies

The formulation derived in Section 2.3 is a nonlinear mixed-integer programming problem. However, this formulation can be linearized through standard techniques from operations research [26], resulting in a mixed-integer linear program (MILP). As such, in theory the MILP formulation could be solved using a number of commercial software packages (e.g., `CPLEX` [27], `LINDO` [28], etc.). It can be shown that this optimization problem is *NP-Hard*, as a variant of the vehicle routing problem with split deliveries and time windows [29, 30, 31], which necessitate heuristic strategies for finding good-quality solutions efficiently [32, 33] when the problem instances are of sufficient size.

The heuristic developed is split into two phases, an Information Routing phase, and a Storage Management phase. The Information Routing phase attempts to route all of the IGRs, from generating resource to consuming resource(s), without consideration of storage limitations on the individual resources. The only constraints taken into account in the Information Routing phase are the bandwidth limitations across the sensor network. The Storage Management phase of the heuristic then factors in the storage limitations of each resource to reduce the solution to one that is feasible.

Figures 2 and 1 provide pseudo-code for the Information Routing and Storage Management phases, respectively. The Information Routing function is input the set of IGRs (and all information about the IGRs), the initial allocation of bandwidth across the network over time (InitCommBW), the minimum percentage (in size) an IGR is able to be reduced and still be useful (MinSizePercent), and the ratio by which to reduce the size of the IGR in sending over the network, if needed (GranularityRatio). Lines 1 to 5 initialize internal variables. Lines 6 to 32 get executed while there are still IGRs on the list to be routed. For each IGR still on the CurListIGRs list, lines 7 to 25 attempt to determine a route for the IGR that allows for at least the minimum size for the IGR to be routed. Line 8 computes the size of the current IGR and line 9 computes the minimal allowable size of the IGR to be routed. While a feasible route for the current IGR has not been found, and the current size of the IGR is larger then the minimum allowable size, lines 11 to 19 get executed. Line 11 calls the function `ComputeRoute`, given the current size of the IGR and the current allocation of bandwidth across the network. Returned from this function are the route for this IGR from generating resource to consumer resource, as well as the time the IGR will completely reach the consumer resource. If the time by which the consumer completely receives the IGR is less than the due date of the IGR, then a potential route for this IGR is found, and the potential time delay for this IGR is computed. Otherwise, the size of the IGR is reduced by the granularity ratio. If a potential route is not found in lines 11 to 19, then this IGR is moved to the NotRoutedIGRs list and removed from the current list of IGRs (lines 21 and 22). Line 23 also sets this IGR to have a delay time of $+\infty$. After the code in lines 7 to 25 is completed, the heuristic greedily selects the IGR with the current minimal potential delay time (line 26). This IGR is removed from the list of current IGRs, and added to the list of routed IGRs. The potential route computed for this IGR is

```
procedure Storage Management(ResourceList, RoutedIGRs, NotRoutedIGRs, MaxStorageSize)
1     CurrentRL ← ResourceList;
2     while CurrentRL ≠ ∅ do
3           CurR ← Choose(CurrentRL);
4           CurrentRL ← CurrentRL \ CurR;
5           for k = 0 to 𝒦 do
6                 for each IGR_ℓ completely transited through CurR by time-step k do
7                       CurR ← RemoveFromStorage(CurR, IGR_ℓ, k, 𝒦);
8                 end for;
9                 for each IGR_ℓ completely consumed by CurR by time-step k do
10                      CurR ← RemoveFromStorage(CurR, IGR_ℓ, k, 𝒦);
11                end for;
12                if Storage(CurR, k) > MaxStorage then
13                      for each IGR_ℓ generated by CurR and completely transited by time-step k do
14                            CurR ← RemoveFromStorage(CurR, IGR_ℓ, k, 𝒦);
15                      end for;
16                end if;
17                ListIGRs ← GetIGRs(CurR, k);
18                SortIGRs ← Sort(ListIGRs);
19                while Storage(CurR, k) > MaxStorage do
20                      IGR_ℓ ← Choose(SortIGRs);
21                      RemoveFromRouting(IGR_ℓ);
22                      NotRoutedIGRs ← NotRoutedIGRs ∪ {IGR_ℓ};
23                      RoutedIGRs ← RoutedIGRs \ {IGR_ℓ};
24                end while;
25          end for;
26    end while;
27    return(RoutedIGRs, NotRoutedIGRs);
end Storage Management;
```

Figure 1: Pseudo-code for the Storage Management phase of the heuristic.

Table 2: Categories, and values used for the number of resources and density over the numerical experiments.

Category	N	Density
Extra Small	3	
Small	4, 5, 6	1, 0.67, 0.33
Medium	8, 10, 12	
Large	15, 20, 25	

saved into the set IGRRoutes in line 29, and the current bandwidth allocation across the temporal network is updated (due to this IGR having its route finalized) in line 30. This process in lines 6 through 32 continues until the set CurListIGRs is empty. The output from the Information Routing function is the set of IGRs that have been routed (along with all of the necessary details of the routes), as well as the set of IGRs not able to be routed.

The Storage Management function takes as input the list of resources, the list of routed IGRs, the list of IGRs that are not able to be routed, and the maximum storage size for each resource. While not all of the resources have been considered, lines 2 through 26 are executed. Line 3 chooses a resource from those still to be considered. Lines 5 to 25 get executed for each time-step of the time horizon. In lines 6 to 8, each IGR that has completely transited through the current resource by the time-step under consideration, gets removed from the resources' storage, from the time-step under consideration to the end of the time horizon. In lines 9 to 11, each IGR that has been completely consumed by the current resource, by the time-step under consideration, gets removed from the resources' storage, from the time-step under consideration to the end of the time horizon. Now, if the storage for the current resource at the time-step under consideration is above the capacity, in lines 13 to 15 each IGR that was generated by the current resource and has been completely transited through the current resource (i.e., completely sent to the next resource along the path to the consumer resource) by the time-step under consideration gets removed from the resources' storage, from the time-step under consideration to the end of the time horizon. Up to this point, all IGRs that were to be routed (as determined from the Information Routing function) are still valid. Line 17 gets the list of IGRs that are scheduled to be generated by the current resource no later than the time-step under consideration, and line 18 sorts this list. While the storage for the current resource at the time-step under consideration is above the maximum allowed, an IGR from the sorted IGR list is chosen, and it is removed from routing. This continues until the storage constraints are met for the current resource at the time-step under consideration. Output from the Storage Management function is an updated list of IGRs chosen to be routed and IGRs not able to be routed.

```
procedure Information Routing(IGRs, InitCommBW, MinSizePercent, GranularityRatio)
1    CurListIGRs ← IGRs;
2    CurCommBW ← InitCommBW;
3    RoutedIGRs ← ∅;
4    NotRoutedIGRs ← ∅;
5    IGRRoutes ← ∅;
6    while CurListIGRs ≠ ∅ do
7        for IGR_j ∈ CurListIGRs do
8            CurSize ← Size(IGR_j);
9            MinSize ← MinSizePercent · CurSize;
10           FoundRoute ← false;
11           while CurSize > MinSize And FoundRoute = false do
12               [RouteIGR_j, ConsumeTime_j] ← ComputeRoute(IGR_j, CurSize, CurCommBW);
13               if ConsumeTime_j < DueTime(IGR_j) then
14                   FoundRoute ← true;
15                   DelayTime_j ← ConsumeTime_j − g_j;
16               else
17                   CurSize ← GranularityRatio · CurSize;
18               end if;
19           end while;
20           if FoundRoute = false then
21               NotRoutedIGRs ← NotRoutedIGRs ∪ {IGR_j};
22               CurListIGRs ← CurListIGRs \ {IGR_j};
23               DelayTime_j ← +∞;
24           end if;
25       end for;
26       j* ← ArgMin(DelayTime);
27       CurListIGRs ← CurListIGRs \ {IGR_j*};
28       RoutedIGRs ← RoutedIGRs ∪ {IGR_j*};
29       IGRRoutes ← IGRRoutes ∪ {RouteIGR_j*};
30       CurCommBW ← Update(CurCommBW, RouteIGR_j*);
31   end while;
32   return(RoutedIGRs, NotRoutedIGRs);
end Information Routing;
```

Figure 2: Pseudo-code for the Information Routing phase of the heuristic.

4 Computational Experiments

4.1 Test Environment

Java jdk 1.8.0_25 was used to code up the heuristic, the actual exact MILP, and the generation of scenario data. CPLEX Optimization Studio V12.6.3 [27] was called to solve the exact MILP. `Matlab R2011b (7.13.0.564)` [34] was used to generate combinations of input parameters, as well as to analyze the results of all experiments. All experiments were conducted on a Dell Precision Tower 7810 with an Intel(R) Xeon(R) CPU E5-2630 v3 @2.40GHz 2.40 GHz with 32.0 GB memory.

4.2 Experimental Results

The aim of our experiments were to investigate the solution quality found by the heuristic, as compared with commercial software, as well as to determine how close to optimal were the solutions found by the heuristic. We used the commercial software CPLEX [27] to solve the linearized formulation of the mathematical model presented in Section 2. However, as can be seen from the results that follow, for some of the scenarios CPLEX was not able to find solutions significantly close to optimal, within the time-limits to find a solution. Thus, we also ran CPLEX to solve the linearized formulation, providing CPLEX an initial solution found from running the heuristic. These two sets of experiments involving CPLEX are described as ''CPLEX w/o Init. Soln.'' and ''CPLEX w/ Init. Soln.'', respectively. To test the heuristic, as well as CPLEX, we created numerous scenarios, varying the number of resources (N) and the density of the underlying communication network topology, as displayed in Table 2.

The maximum density of the underlying communication network topology is given by Equation (33), where N is the set of nodes, or resources, of the network and E is the set of communication links, or edges, between nodes of the network. So, a maximum density of 0.8 means that 20% of the edges have been deleted from the complete network. We note that deleting these edges has the same effect as enforcing that those edges have a maximum bandwidth of 0 over the time-horizon. Also it is important to point out that this gives the maximum density of the network over any given time. In actuality, based on temporal bandwidth parameters defined ('Bandwidth per Directed Edge per Resource' in Table 3), there exist the potential for two nodes to have 0 available direct bandwidth at a given time-step, even though this edge was not explicitly removed from the network. This can be seen as modeling the possibility for a node to be kinematically out of range of another, and hence not able to communicate, or for a node to be operating in 'radio silence' at that

Table 3: Parameter categories and values used for the numerical experiments. For those parameters with only a minimum value, that is the fixed value for the parameter. For those parameters that have a minimum and maximum value, the distribution listed is the probability distribution through which a specific parameter is determined, via a random draw.

Parameter	Min. Value	Max. Value	Dist.
Number Time-Steps	20		
Bandwidth per Directed Edge per Resource	0	100	Uniform
Number of IGRs in Scenario per Resource	5	10	Uniform
Time Delay Between IGR Generation and Consumption	7	20	Uniform
Resources that Neither Generate nor Consume IGRs	$\lceil 0.2 \cdot N \rceil$		
IGR Size	25	250	Uniform
Storage Capacity per Resource	750	1500	Uniform
Granularity Ratio	0.9		
MinSizePercent	0.8		

Table 4: Solution quality analysis, according to Equation (34), for the mathematical formulation using CPLEX, without and with an initial input solution.

		CPLEX w/o Init. Soln.			CPLEX w/ Init. Soln.	
Num. Resources	Density	% Time Opt. Soln. Found	Mean Soln. GAP	% Time Feasible Soln. Not Found	% Time Opt. Soln. Found	Mean Soln. GAP
3	1	80	0.0012	0	80	0.0012
3	0.67	100		0	100	
3	0.33	100		0	100	
4	1	100		0	100	
4	0.67	90	0.0013	0	90	0.0012
4	0.33	90	0.0001	0	90	0.0001
5	1	80	0.0002	0	90	0.0001
5	0.67	60	0.0029	0	50	0.0027
5	0.33	70	0.0006	0	70	0.0006
6	1	60	0.0207	0	50	0.0035
6	0.67	60	0.0022	0	70	0.0093
6	0.33	100		0	100	
8	1	0	0.0343	0	0	0.0133
8	0.67	0	0.0775	0	10	0.0368
8	0.33	20	0.0109	0	20	0.0115
10	1	0	0.1149	0	0	0.0274
10	0.67	0	0.1031	0	0	0.0240
10	0.33	0	0.0636	0	0	0.0890
12	1	0	0.2145	0	0	0.0390
12	0.67	0	0.1308	0	0	0.0349
12	0.33	0	0.3658	0	0	0.2128
15	1	0	0.4977	0	0	0.0678
15	0.67	0	0.2130	0	0	0.0637
15	0.33	0	0.4390	10	0	0.1810
20	1	0	0.9583	90	0	0.2988
20	0.67	0	0.6205	30	0	0.1122
20	0.33	0	0.2709	10	0	0.2507
25	1	0	1.0000	100	0	0.2614
25	0.67	0	1.0000	100	0	0.3739
25	0.33	0	0.3702	10	0	0.2155

time-step, due to external environmental conditions (e.g., weather, topography, etc). In addition, this can model a node failure [35, 36, 37] if a particular node has 0 available bandwidth at a given time over all edges

Table 5: Average normalized distance, computed according to Equation (35), between the heuristic and CPLEX with/without the initial solution, for those cases where CPLEX found at least a feasible solution. Note: An entry is left blank whenever CPLEX was not able to find a feasible solution for any of the scenarios.

Num. Resources	Density	CPLEX w/out Init Soln	CPLEX w/ Init Soln
3	1	-0.132	-0.132
3	0.67	-0.127	-0.127
3	0.33	-0.115	-0.115
4	1	-0.072	-0.072
4	0.67	-0.174	-0.174
4	0.33	-0.330	-0.330
5	1	-0.098	-0.098
5	0.67	-0.194	-0.194
5	0.33	-0.212	-0.212
6	1	-0.071	-0.089
6	0.67	-0.151	-0.145
6	0.33	-0.481	-0.481
8	1	-0.064	-0.085
8	0.67	-0.079	-0.128
8	0.33	-0.400	-0.400
10	1	0.024	-0.071
10	0.67	-0.013	-0.120
10	0.33	-0.281	-0.254
12	1	0.158	-0.066
12	0.67	0.070	-0.100
12	0.33	0.300	-0.102
15	1	0.779	-0.051
15	0.67	0.090	-0.086
15	0.33	0.491	-0.114
20	1	1.186	-0.037
20	0.67	0.536	-0.066
20	0.33	-0.068	-0.108
25	1		-0.035
25	0.67		-0.036
25	0.33	0.109	-0.104

connecting it to other nodes.

$$\text{Density} = \frac{|E|}{|N| \cdot (|N| - 1)} \tag{33}$$

For each combination of number of resources and density, 10 Monte-Carlo scenarios were created randomly. For each scenario generated, the remaining parameters were derived from the information in Table 3, and then kept fixed for the scenario. In addition, the coefficients of the two terms in the objective function (time-delay and size), ω_d and ω_s, were fixed to 0.7 and 0.3 for all scenarios. Each approach (''CPLEX w/o Init. Soln.'', ''CPLEX w/ Init. Soln.'', and the heuristic) were given 900 seconds to solve each scenario instance. For those cases where CPLEX did not solve the problem to optimality within the 900 seconds, the best solution found by CPLEX is considered. We note that there are cases in the tables to follow where CPLEX was not able to find even a feasible solution by the time-limit, hinting at the limitations of commercial software as the size of the problems get large.

Table 4 examines the results just from the exact approach, using CPLEX (without and with an initial solution). For each combination of number of resources and communication network density, columns 3–5 deal with the solution found by CPLEX without an initial solution, while columns 6–7 are concerned with the solution found by CPLEX with an initial solution provided. Columns 3 and 6 show that as the problem size gets larger, CPLEX has a more difficult time in finding the optimal solution. And for most of the problem sizes, in terms of providing an optimal solution, there is no difference between CPLEX without or with an initial solution. Column 5 shows the percentage of times CPLEX without an initial solution is unable to find a feasible solution to the problem. There is always the trivial feasible solution where no IGRs are routed, so CPLEX unable to find a feasible solution means that CPLEX ran out of memory in the creation of the mathematical model. Again this is as expected; as the problems get larger, more variables and constraints are needed, and eventually memory issues are encountered. Columns 4 and 7 show the mean gap between the upper bound on the solution value computed in CPLEX and the best solution found by CPLEX. For each scenario, this gap is computed as in Equation

Table 6: Mean Time Delay between generation and consumption of IGRs chosen to be routed, for each solution approach. For those cases where CPLEX is not able to find a feasible solution within the time-limit for any scenarios, the entry is left blank. The results are shown as mean:standard deviation.

Num. Resources	Density	CPLEX w/out Init Soln	Heuristic	CPLEX w/ Init Soln
3	1	8.956:0.758	5.070:1.120	8.937:0.738
3	0.67	9.691:1.387	4.968:0.897	9.658:1.389
3	0.33	10.223:0.993	4.651:0.722	10.194:0.960
4	1	9.268:0.954	4.105:0.335	9.278:0.952
4	0.67	9.867:0.734	5.059:0.680	9.856:0.697
4	0.33	10.096:1.153	6.036:1.208	10.128:1.168
5	1	9.004:0.676	3.773:0.327	9.030:0.663
5	0.67	9.762:0.779	5.073:0.433	9.735:0.756
5	0.33	9.588:1.137	5.321:0.700	9.595:1.128
6	1	9.300:0.571	4.021:0.296	9.270:0.582
6	0.67	9.502:0.988	5.257:0.345	9.571:0.953
6	0.33	9.925:1.093	6.361:0.861	9.944:1.050
8	1	9.201:0.700	4.016:0.247	9.186:0.639
8	0.67	9.233:0.558	4.704:0.340	9.426:0.830
8	0.33	10.047:0.665	6.433:0.560	10.099:0.667
10	1	9.010:0.560	3.807:0.321	9.248:0.613
10	0.67	8.504:0.580	4.517:0.316	8.788:0.658
10	0.33	9.460:0.457	6.117:0.519	9.928:0.984
12	1	8.632:0.482	3.846:0.234	9.242:0.494
12	0.67	8.950:0.458	4.684:0.313	9.324:0.497
12	0.33	9.362:0.603	6.194:0.334	11.301:0.979
15	1	7.573:0.398	3.820:0.122	9.599:0.288
15	0.67	8.590:0.698	4.411:0.204	9.197:0.570
15	0.33	9.449:1.149	5.924:0.569	11.035:0.874
20	1	7.633:0.000	3.688:0.148	9.692:0.348
20	0.67	7.586:0.312	4.376:0.068	9.961:0.541
20	0.33	8.684:0.512	5.866:0.269	10.534:1.066
25	1		3.721:0.096	9.629:0.312
25	0.67		4.366:0.211	10.469:0.287
25	0.33	8.471:0.424	5.537:0.139	10.359:1.046

(34), where ub is the upper bound on the solution value and sf is the solution found by CPLEX. We note that when CPLEX finds the optimal solution, $ub = sf$ and the GAP is therefore 0, and when CPLEX is not able to find even a feasible solution, we assigned the trivial solution of $sf = 0$ (the worst feasible solution for the model), resulting in a GAP value of 1.

$$GAP = \frac{ub - sf}{ub} \tag{34}$$

What is apparent from Columns 4 and 7 of Table 4 is that CPLEX with an initial solution performs better most of the time as compared to CPLEX without an initial solution. In most cases, there is a significant decrease in the mean gap. However, there are a few cases that are counter-intuitive. Specifically when the number of resources is 6, 8, or 10 and the densities are 0.67, 0.33, or 0.33 respectively. For these three combinations, CPLEX without an initial solution has a smaller mean gap than does CPLEX with an initial solution. This is because there was one scenario in each of these combinations where the heuristic found a solution not close to the optimal, and CPLEX had a difficult time given the heuristic solution as the initial solution, i.e., the heuristic found a local maximum, and CPLEX had a difficult time finding a solution better than this local maximum.

Table 5 shows the average normalized distances between the heuristic solutions and the CPLEX solutions without and with the initial solution. Each normalized distance is computed according to Equation (35), where $\{x_j^h, s_j^h\}$ is the solution found by the heuristic, $\{x_j^c, s_j^c\}$ is the solution to the linearized formulation using CPLEX, and f is the objective function. Since we have a maximization problem, values of Equation (35) larger than 0 indicate that the heuristic is finding a better solution than is CPLEX, while those less than 0 indicate an heuristic solution worse than that found by CPLEX. First off, as expected, all of the values in the last column of Table 5 are negative, because the worst solution 'CPLEX with initial solution' can find is the same solution as found by the heuristic since this is the solution provided as input to CPLEX. For less than 12 resources, CPLEX without an initial solution

Table 7: Average runtime (in seconds) for each solution approach to find a solution. For those cases where CPLEX is not able to find a feasible solution within the time-limit, the returned time by CPLEX is still used in this computation.

Num. Resources	Density	CPLEX w/out Init Soln	Heuristic	CPLEX w/ Init Soln
3	1	218.705	0.002	224.092
3	0.67	48.794	0.000	132.075
3	0.33	10.150	0.000	12.552
4	1	22.205	0.003	30.301
4	0.67	179.300	0.000	190.852
4	0.33	123.936	0.001	155.227
5	1	388.937	0.000	337.506
5	0.67	433.410	0.002	498.094
5	0.33	465.390	0.001	444.524
6	1	469.135	0.002	634.885
6	0.67	595.497	0.002	502.483
6	0.33	109.626	0.000	161.185
8	1	901.857	0.008	901.757
8	0.67	901.863	0.006	865.321
8	0.33	822.245	0.005	761.825
10	1	904.283	0.006	903.338
10	0.67	905.129	0.005	905.198
10	0.33	903.219	0.002	903.443
12	1	904.817	0.011	905.141
12	0.67	905.752	0.003	906.519
12	0.33	905.980	0.003	904.835
15	1	910.124	0.044	910.081
15	0.67	1310.217	0.005	1071.725
15	0.33	1161.690	0.005	1124.190
20	1	2872.113	0.083	2066.906
20	0.67	1283.765	0.009	994.237
20	0.33	958.069	0.005	1153.269
25	1	4612.964	0.128	2845.093
25	0.67	6458.231	0.006	6317.989
25	0.33	1672.833	0.006	1980.635

is also able to find a better solution on average than is the heuristic. However, after this point, for almost all cases the heuristic is finding a better solution than is CPLEX without an initial solution. Again, this is to be expected, because as the size of the problem grows, commercial software will have more difficulty in solving the larger problems and the heuristic will begin to find better solutions. But in comparing the heuristic with CPLEX with an initial solution, the average normalized distance between solutions is never that large on average, showing that the heuristic is able to find a good-quality solution as compared with commercial software.

$$\frac{f\left(x_j^h, s_j^h\right) - f\left(x_j^c, s_j^c\right)}{f\left(x_j^c, s_j^c\right)} \tag{35}$$

Table 6 presents, for those IGRs chosen to be routed, the mean time delay between IGR generation and IGR consumption. Column 3 of Table 6 shows the mean delay for the CPLEX solution when no initial solution is provided, Column 4 shows the mean delay for the heuristic solution, while Column 5 shows the mean delay for the CPLEX solution when an initial solution is provided. As is clear from this table, for those IGRs routed, the heuristic is able to route them much quicker than is CPLEX. This is regardless of the size of the problem. The rationale for this is that the heuristic is choosing IGRs to route in a greedy way, while CPLEX considers the routing of all IGRs simultanously.

Table 7 presents the average runtime for each of the approaches, in seconds. As is clear, the developed heuristic is quite fast in finding solutions for all of the scenario classes. It is also clear that the heuristic takes longer to run on average when the communication network topology is dense as opposed to sparse. Both of the CPLEX approaches take much longer than does the heuristic. Even when the number of resources is as small as 8, these approaches reach (or almost reach) the time limit of 900 seconds. Hence, to get sightly better solutions (according to the objective function) CPLEX needs much more time than does the heuristic. We note that while CPLEX was given a time limit of 900 seconds, there were cases where CPLEX does take

longer to complete. This is due to CPLEX not being able to 'halt' processing exactly when the time limit is reached.

5 Conclusions and Future Research

In this research, we have examined the problem of routing information amongst a set of resources, when there exists a dynamic communication network topology with limited, time-varying, bandwidth, over a given time horizon. A rigorous mathematical formulation was developed to model the problem, with the objective being to minimize the time delay of the information that can be routed while at the same time maximizing the size of the information routed. This formulation is nonlinear, but a linearized version can be derived through standard techniques of operations research. As the problem is NP-hard, a heuristic has been developed to efficiently find good-quality solutions. Numerous Monte-Carlo simulations were performed on problems with varied resources and communication network topology density. As can be seen in the results, for smaller sized problems commercial software is able to find a better solution on average than is the heuristic. And when the commercial software uses as input the solution found by the heuristic, even better results are obtained. As the size of the problem increases, however, the heuristic begins finding much better solutions compared with the commercial software. And even when the commercial software is provided the heuristic solution as input, the commercial software is not able to improves significantly on this solution. Looking at the metric of the actual time-delay for those IGRs chosen to be routed, it is clear that the heuristic is better able to route those chosen IGRs. When coupled with the time needed for the heuristic to find a good quality solution versus the commercial software, it is clear that the heuristic is outperforming the commercial software.

Future research includes looking at specific network topology structures from realistic military and civil applications, considering the related problem of finding the minimum network temporal connectivity necessary to ensure certain information is able to be routed within time bounds, as well as considering the problem from a decentralized control paradigm, when no resource has knowledge about the network as a whole, but rather must consider only local network knowledge and make independent routing decision. In addition, incorporating the concept of network packet errors will provide added complexity and realism to the problem and scenarios considered.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This research was Supported by Office of Naval Research Project N00014-15-C-5163.

References

[1] M. J. Hirsch, A. Sadeghnejad, and H. Ortiz-Pena, "Shortest paths for routing information over temporally dynamic communication networks," in *Proceedings of the IEEE Military Communications Conference*, pp. 587–591, 2017.

[2] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: A survey," *Computer Networks*, vol. 38, pp. 393–422, 2002.

[3] J. Al-Karaki, , and A. Kamal, "Efficient virtual-backbone routing in mobile ad hoc networks," in *Elsevier Computer Networks*, vol. 52, pp. 327–350, 2008.

[4] E. Huang, W. Hu, J. Crowcroft, and I. Wassell, "Towards commercial mobile ad hoc network applications: a radio dispatch system," *6th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pp. 355–365, 2005.

[5] "Army 18.1 Small Business Innovation Research (SBIR) Proposal Topic 025: An Adaptively Covert, High Capacity RF Communications / Control Link," tech. rep., 2018.

[6] "Air Force 18.1 Small Business Innovation Research (SBIR) Proposal Topic 019: Novel Battle Damage Assessment Using Snesor Networks," tech. rep., 2018.

[7] C. Howard, "UAV command, control, and communication," *Military and Aerospace Electronics*, vol. 24, 2013.

[8] B. Vignac, F. Vanderbeck, and B. Jaumard, "Reformulation and decomposition approached for traffic routing in optical networks," *Networks*, vol. 67, pp. 277–298, 2016.

[9] J. Brodkin, "Tropical storm Harvey takes out 911 centers, cell towers, and cable networks," *ars Technica*, August 28 2017.

[10] L. K. Comfort and T. W. Haase, "Communication, choerence, and collective action: The impage of Hurrican Katrina on the communication infrastructure," *Public Works Management & Policy*, vol. 11, no. 1, pp. 1–16, 2006.

[11] U.S. Air Force, "USAF Strategic Master Plan," tech. rep., 2015.

[12] "Air Force 17.1 Small Business Innovation Research (SBIR) Proposal Topic 047: Resilient Communication for Contested Autonomous Manned/Unmanned Teaming," tech. rep., 2017.

[13] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," in *Proceedings of the IEEE*, vol. 85-1, pp. 6–23, 1997.

[14] E. Bosse, J. Roy, and S. Wark, eds., *Concepts, Models, and Tools for Information Fusion*. Artech House, 2007.

[15] F. A. A. U.S. Department of Transportation, "Aviation Maintenance Technician Handbook – Airframe, Volume 2," tech. rep., 2012.

[16] "Navy 17.A Small Business Technology Transfer (STTR) Proposal Topic T029: Multi-Vehicle Collaboration with Minimal Communication and Minimal Energy," tech. rep., 2017.

[17] I. Jebadurai, E. Rajsingh, and G. Paulraj, "Enhanced dynamic source routing protocol for detection and prevention of sinkhole attack in mobile ad hoc networks," *International Journal of Network Science*, vol. 1, pp. 63–79, 2016.

[18] U. Paul, M. Buddhikot, and S. Das, "Opportunistic traffic scheduling in cellular data networks," *IEEE International Symposium on Dynamic Spectrum Access Networks*, pp. 339–349, 2012.

[19] J. Al-Karaki, R. Ul-Mustafa, and A. Kamal, "Data aggregation and routing in wireless sensor networks: Optimal and heuristic algorithms," in *Elsevier Computer Networks*, vol. 53, pp. 945–960, 2009.

[20] M. J. Hirsch, H. Ortiz-Pena, R. Nagi, A. Stotz, and M. Sudit, "On the optimization of information workflow," in *Dynamics of Information Systems: Mathematical Foundations* (A. Sorokin, R. Murphey, M. Thai, and P. Pardalos, eds.), vol. 20, pp. 43–65, Springer, 2012.

[21] M. J. Hirsch and H. Ortiz-Pena, "Information supply chain optimization with bandwidth limitations," *International Transactions in Operational Research*, vol. 24, no. 5, pp. 993–1022, 2017.

[22] E. Grotli and T. Johansen, "Path and data transmission planning for cooperating uavs in delay tolerant network," in *Globecom Workshop*, pp. 1568–1573, IEEE, 2012.

[23] S. Basagni, C. Petrioli, R. Petroccia, and M. Stojanovic, "Optimizing network performance through packet fragmentation in multi-hop underwater communications," in *OCEANS*, pp. 1–7, IEEE, 2010.

[24] P. Kansal and A. Bose, "Bandwidth and latency requirements for smart transmission grid applications," *IEEE Transactions on Smart Grid*, vol. 3, pp. 1344–1352, 2012.

[25] M. Greenberg, *Advanced Engineering Mathematics*. Pearson, 2nd ed., 1998.

[26] L. Wolsey, *Integer Programming*. Wiley, 1998.

[27] I. CPLEX, "http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/," Accessed January 2017.

[28] LINDO, "http://www.lindo.com/," Accessed July 2016.

[29] J. Lenstra and A. R. Kan, "Complexity of vehicle and scheduling problems," *Networks*, vol. 11, pp. 221–227, 1981.

[30] M. Solomon and J. Desrosiers, "Time window constrained routing and scheduling problem," *Transportation Science*, vol. 22, pp. 1–13, 1988.

[31] M. Dror and P. Trudeau, "Split delivery routing," *Naval Research Logistics*, vol. 37, pp. 383–402, 1990.

[32] G. Ausiello, M. Protasi, A. Marchetti-Spaccamela, G. Gambosi, P. Crescenzi, and V. Kann, *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer-Verlag New York, Inc., 1st ed., 1999.

[33] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA: W. H. Freeman & Co., 1979.

[34] Matlab, "http://www.mathworks.com/products/matlab/," Accessed January 2017.

[35] B. Szymanski, X. Lin, A. Asztalos, and S. Sreenivasan, "Failure dynamics of the global risk network," *Scientific Reports*, vol. 5, pp. 1–14, 2015.

[36] R. Liu, M. Li, and C. Jia, "Cascading failures in coupled networks: The critical role of node-coupling strength across networks," *Scientific Reports*, vol. 6, pp. 1–6, 2016.

[37] K. Kim, J. Jin, and S. Jin, "Classification between failed nodes and left nodes in mobile asset tracking systems," *Sensors*, vol. 16, pp. 1–18, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 328-338 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Q-Learning versus SVM Study for Green Context-Aware Multimodal ITS Stations

Adel Mounir Said[1], Emad Abd-Elrahman[*,1], Hossam Afifi[2]

[1]*National Telecommunication Institute , Cairo, Zip Code 11768, Egypt.*

[2] *Telecom SudParis, IMT Saclay, Zip Code 91120, France.*

ARTICLEINFO

Article history:
Received: 15 August, 2018
Accepted: 07 October, 2018
Online: 16 October, 2018,

Keywords:
Q-Learning
SVM
Context Awareness
Multimodal ITS

ABSTRACT

Intelligent Transportation Systems (ITS) applications can take big advantage of Context Awareness approaches. Parameters such as user mobility, passengers comfort reaction and pollution emission levels (CO_2) can enrich such applications during the decision making phase. Moreover, the expanding in ITS services offers great opportunities for travelers to find the best route to reach their destinations with the lowest or fair costs. It can offer a selecting methodology for optimal route that adapted with some processing parameters like CO_2 level, ticket cost, waiting or connection times and the overall traveling time plus the comfortability reaction for each means of transportation) in real time environment using Machine Learning (ML) tools like Q-Learning or SVM: Support Vector Machines. This paper aims at conducting a comparison study for green ITS routes (i.e. the lowest CO_2 levels). The study compares between Q-Learning and SVM techniques for identifying different variety of routes between two stops as ranked routes from best to lowest based on some traces gathered from some known transportation traces. Reinforcement Q-Learning is applied to validate the first phase in our approach to recommend the best means and SVM is used to validate the prediction phase about the best route among different routes built based on three means of transportation (metro, train and bus).

1 Introduction

The ITS network is considered as the most important element of concern to the countries in the modern smart cities era. It is being the arteries that pass through continues chain of business activities, social, cultural and tourism that promote the march of the national economy for its role in providing services to the sectors of production and other services, and is reflected from providing employment opportunities. Therefore, the new ITS construction projects occupy the first initiative in sustainable development programs and projects to achieve higher rates of growth and development through the provision of passenger traffic between all stations or cities of the proposed new development. The governments directions are focused on planning and implementations of future stations that help in the preparations and developments of the countries through intelligent decision mechanisms.

Today, the need for an ITS solution becomes crucial to provide a smart solution for managing the traffic on country's roads network. Significant outcomes expected of using ITS are to provide a smooth, smart, and sustainable transport means.

Moreover, distributing the passengers over all the available transportation means (like buses, cars, trains, planes, and ships) can lead to some kind of balance between them and help in solving the problems of traffic jams. This aim can be achieved if there will be an implementation of intelligent systems having clear information about some states statistics like current traffic means available, roads status, passengers per day and pollution's emitted per trip. Those implemented systems can predict the long term situation of city or state transportation system behavior and recommend the future planning for new smart cities. Another important thing is the world climate changes due

*Corresponding Author: Emad Abd-Elrahman, Address: 5 Mahmoud El Miligui Street, 6th district - Nasr City, Cairo - Egypt, Phone: 0020-100720268 & Email: emad.abd_elrahman @ telecom-sudparis.eu

https://dx.doi.org/10.25046/aj030539

to transportation activities. Those changes are serious and can be affected by the huge carbon emission levels from different means of transportation and hence the dangerous consequences on the humanity life. This could be an incentive to optimize and reduce the CO_2 emission either from public or private transportation means that will lead the concept of Green-ITS.

Context Awareness (CA) approach can be used to enrich the ITS services as it can adapt the control system dynamics. This shall grasp the potential advantages, which influence the control strategies taken to manage the transportation system especially, in multimodal traffic management. The multimodal traffic station has different crossings of different transportation means like metro,train, bus, and etc.

This paper is an extension of the work originally presented in (ICECTA) conference [1].

Through the work, we improved the previous implementations about advanced ITS traffic flows management [2]. The aim of the previous paper was to choose among different transportation means the best one. We considered the recommendation of the best transport mean from the available three means of transportation in multimodal station with crossing of metro, bus and train. This recommendation is done based on some user and system parameters at any time (t) for any passenger decides to reach a specific destination using the previous means. We used a reward based Q-learning approach to choose the best transport means available in multimodal stations. In the second phase of this work, we propose using Support Vector Machines (SVM) for their reputation in classification accuracy or prediction speed comparing to the previous approach.

The context of SVM foundations had been developed in 1995 by Vapnik [3] and obtained high popularity in different applications due to some attractive and intelligent features, and the promising performance in empirical tests [4].

To conclude our work, the goal is to select among different routes available between two stations the best one for the destination. The selection is carried out by classifying some collected ITS traces about all the possible routes including combinations between different means of transportation at any time (t). Moreover, we considered the influence of new parameters that are missed in the previous work [2] like:

- The passenger waiting time at the starting station till the arrival of transportation mean.
- The passenger waiting time if there is a connection(s) between different or same transport means to reach specific destination.
- The measurement of CO_2 pollution from each trip.

Furthermore, the performance evaluation study in this work is done based on real traffic statistics that gathered in real time from *SNCF Transilien* [5] and *Vianavigo* [6] French sites. This data covered some working and week-end days during one month testing period from the French transportation network of Paris area.

The construction for paper is structured as follows: In Section 2, we introduce the related work done in ITS decision-makings analysis and management using Machine Learning (ML) tools. Then, Section 3 presents the Reinforcement Q-Learning model then the SVM for the context aware ITS multimodal stations. Afterward, the evaluations done for the proposed models and their results are compared with different selected use cases of working and week-ends days in Section 4. This is conducted through extensive simulations based on collected real-time statistics from the French *SNCF Transilien* site [5]. Finally, Section 5 concludes and highlights the perspectives of this work.

2 Work Background

An Intelligent Agent (IA) term in ITS domain has been defined as: the agent which receives precepts from the environment and based on that takes smart actions [7]. The IA is considered as the transportation station in our study. This agent has the capability to take decisions based on the gathered reactions from the passengers. Upon received those reactions, a context-aware learning system is used to optimize the taken decisions. Then, it proposes an accurate guides for all passengers crossing this IA (i.e. the multimodal station). Hereafter, we highlight some relevant insights that tackled the intelligent transportation prediction systems based on reported reactions.

The following literature is organized in two sections. The first one reviews some propositions based on using reinforcement and Q-learning techniques [8-14]. The second section reviews the propositions based on using SVM techniques in ITS. As it seems to our problem context, there is no related work uses SVM learning algorithm to detect the travel time of each transport mean and hence predict the best one at each time (t). As will be seen, the most of these works in this section is relating to detecting the travel time for bus and cars on highways based on SVM.

2.1 Q-Learning based Prediction Systems

Q-Learning is typically used in reward based reinforcement learning in different ITS domains. The technique of arrival time estimation proposed in [8] a prediction model. This model considered its decision based on common context for both vehicles and their drivers. The proposed solution considered an increasing in the anticipated time of arrival based on some observations in the traffic flow patterns.

Moreover, the pattern anticipates the vehicles arrival time from the history of the passed routes (i.e. the whole previous traveled routes). The main weak point of this model is the lack of consideration either in backup routes to be used as parallel routes or in emergency situations due to traffic jam. But, this proposal

is a good solution for ITS tracking systems.

The work in [9, 10] proposed the reinforcement learning model for the traffic-based control. This is for the case of classical pre-timed systems in signal control and the control agents are also independent. Through the model, the authors implemented the learning system for their control agent based on reactions' rewards as a closed-loop control agent. This agent will interact with the environment to converge with the control policy for learning and achieving an optimal mapping for the optimal required control action and the environments state.

Moreover, the authors had been considered in another work [11] the use of the reinforcement learning approach in a decentralized system. They proposed an adaptive real-time agent for traffic signal control system that ables to minimizing the vehicles total travel time. Based on accumulated reward, Q-Learning agent is used to learn the optimal mapping between the environments state in one side and the corresponding control action in the other side.

The Q-learning model for traffic control considered in [11] is based on time varying and stochastic problems in the traffic flow while in [12], authors investigated the snag of agent-based self-optimization against multiple polices. They proposed the use of distributed w-learning reinforcement learning model. The purpose behind this idea applied in self-organized traffic control systems, is the performance improvement from multiple policies deployments simultaneously.

But, the work in [13] proposed a prediction model for the flow of traffic depending on the multi-agent reinforcement learning. This model is based on multi-cross roads control for the traffic signal.

Finally, authors in [14] concentrated on the coordination between control agents in order to adapt the roads traffic signals. Therefore, they propose to use an independent control agent mode besides an integrated mode merged with reinforcement learning. This model is capable of solving the agents' communication problems. Their results indicate that there is a delay reduction for the connection travel time according to their model proposition.

2.2 Classification Systems based SVM

SVM as a supervised machine learning tool that is based on simple principles, originated from statistical learning theory [3]. The SVM simplicity is coming from the applying of simple linear methods on the data [15].

The accuracy as well as simplicity of the SVM models are the key for many contributions based on classification and regression.

In this part, we present the most research directions that used SVMs in ITS domain as follows:

The first main concern in this direction using SVMs was focusing on the time prediction for the trip. The authors in [16-18] introduced different models based on SVMs to predict the time of each travel on the highway routes. While in [19], the authors proposed their travel time prediction model using the SVM in urban transport networks.

Other contributions considered SVM for the bus travel times prediction. In [20], SVM based hybrid model was introduced for this time prediction. Their technique was based on combining between the classical SVM, the Grubbs test method and an adaptive algorithm.

In [21], a new bus travel time prediction model for multiple dynamic routes is proposed. The model used SVM with Kalman filtering technique.

The authors in [22] proposed their time prediction model for the flow of rail transit passengers in the city of Beijing using SVR to improve the rail transportation management. They used the PLSR: Partial Least Squares Regression method to resolve multi colinearity among the dependent variables.

Last direction in our study focused on the research directions in the traffic signs detection or prediction. In [23-26], the authors developed automatic systems for either traffic sign detection or recognition using SVM. Through [26], the authors proposed the use of SVM as a recognize module to analysis the keep-clear signs to enhance the traffic management in cities.

Another work [27] was focusing on using SVM for automatic video based vehicle detection. This is for an automatic traffic surveillance system based on ITS.

In [28], the authors proposed a design for a long-haul bus. This device is used for driving safety warning based on SVM. It is based on classifying the lane departure and forward collision.

To the best of our knowledge, SVM is not used in any ITS based model to nominate the best traffic mean or route for passengers to reach specific destination at time (t) under the context of multi-modal traffic. In the following section, the proposed model will be described and the both learning algorithms as well.

3 ITS Proposed Models

3.1 Green Transport Decision System Model (GTDS)

Usually, passenger(s) arrive to the Transport Station (TS) for reaching certain destination. Therefore, they have to decide taking a Route i (Ri) which may includes one or more transport mean at time (t). Here, a model for Green Transport Decision System (GTDS) is proposed for facilitating the passengers travels. GTDS model nominates the best Ri for passenger(s) based on Q-learning. The system nominations rely on passengers reward reactions. In contrasting SVM with Q-learning, SVM is faster and has a higher accuracy. Therefore, in this paper SVM is proposed to be used in the second phase of this work. Figure 1 demonstrated the model main inputs and the collected reactions as system feedbacks.

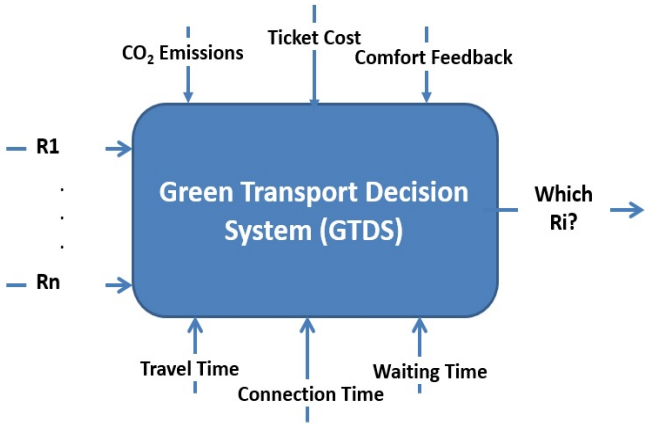


Figure 1: The main inputs and feedbacks for context-aware GTDS framework

Stations apply the GTDS model should consider having multiple routes to the same destination. This is to provide different nominations at different situations aspects. According to that, the proposed model is applied on the departure station called: (*Gare De Paris Montparnasse*) and the arrival station called: (*Gare De Paris Nord*) as a study use case. The model is simulated by different week days activities (i.e. both weekend days and work days). The studied period consists of 25 quarters starting at 7:00 am until 1:00 pm. This period is intended to contain all the expected traffic conditions; either the rush hours or the free from traffic jam situations.

The choice of the best (Ri) is affected by many parameters, the most relevant are determined in the proposed model and illustrated in Fig. 1. The first factor is the passengers travel duration to reach the trend-setting. This factor will be considered taking into account the respect time for each means of transportation shown in Figure 2. The curves demonstrated in this figure have been obtained through the history of each transportation mean by collecting several observations. As indicated, the use of metro is more frequent and comfort than other transport means. The train achieves rank two and finally the bus. For this parameter which reflects respecting the arrival time, we are interested on some statistics which provide a wide sight for how the transport means follow or not the scheduled times arrival. This insight comes from the published SNCF statistics or real time scheduling plan for the transport means on the on-line site [5].

As clear in Figure 2, the best of the transport means in respecting the arrival times is the Metro. Even it may exceed its scheduled arrival times due to the rush hours. Therefore, when a passenger comes to the station and check in to see the best transportation tool to go to his destination D, the decision system inside the station uses the reactions from the station D about previous respecting arrival schedules to update the immediate reward function r when a passenger tries to take a transportation mean.

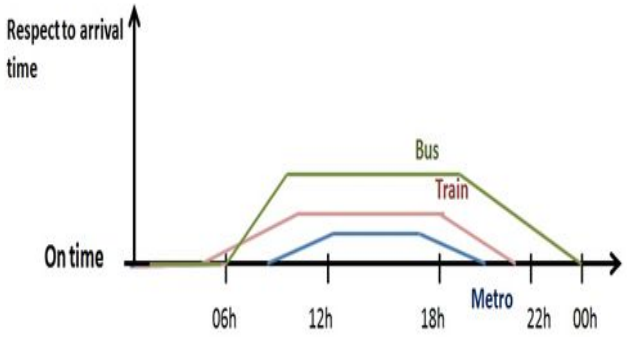


Figure 2: Arrival times respect ratio for the main three transportation means, the metro comes in the first rank followed by local train and then bus as worst respect ratio to arrival on time

By doing so, the decision system firstly tries to optimize its future decisions biased from the previous experience along the day.

Secondly, the travel cost which could affect the system decisions. As known, the most of passengers propose the use of the cheapest transport mean.

Thirdly, if there is a waiting time before catching the transport mean.

Fourthly, in the travels of multiple transport means, the connection time is considered as an effecting parameter.

Fifthly, the comfortability coefficient for the use of Ri at any time (t). This coefficient reflects the passenger satisfaction degree of current experience about using any route Ri. The impact is influenced by the available free chairs in rush hours comparing to the number of passengers intend to use this route. Therefore, passengers are more interested of using the more comfortable transport means. The comfortability reaction of the passengers is shown in Figure 3. The mentioned parameters are function of time even the ticket price. In some countries, ticket cost plan changes along the week days (i.e. working days and off days). At present, the major of the world countries became concerned about the CO_2 emission because of the catastrophic effects on the world climate.

In this contribution, we are interested in adding the CO_2 parameter that affects the selection of the transport means and hence the best route. In France, the government encourages the people to use the public transport means to decrease the emission of the CO_2 and hence the air pollution. Moreover, France, during high air pollution days, encourages the use of public traffic means for free, and prevents the even/odd cars plate numbers in Paris area.

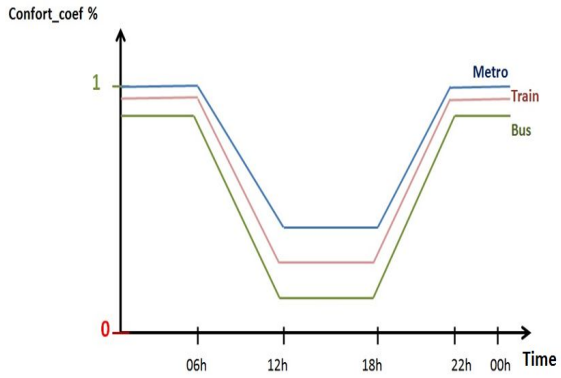


Figure 3: Passengers Comfortability Feedbacks/Reactions, comfort coefficient equal one (comfort_coef=1) represents high satisfaction from the passengers while the low level equal zero for unsatisfied trips

3.2 Reinforcement Learning Model

Using reinforcement learning, any TS can learn from its experience to guide passengers and thus, optimize its interactions for the future decisions. The mentioned experience is stated as rewards and sanctions from the prior taken decisions. This model is initialized at each TS benefiting from the collected passengers reactions. When a passenger uses Ri to reach a destination, this Ri will be rewarded or sanctioned according to these parameters: the travel time for reaching certain destination, the passengers comfortability reaction, and finally the travel cost of every Ri. Consequently, each TS will tune its travel means nominations from its gained experience.

In this paper, the TS decision problem is formulated as a Markov Decision Problem (MDP), that can be solved by the reinforcement learning algorithm. The TS will elects the best Ri for reaching a destination biased by its experience at each day hour, or another duration period based on the traffic. Next to that, the system evaluates itself by assessing the decision taken as a reward or sanction. The MDP is defined as a tuple $[i, t, a, r]$.

- i: stands for the object that is considered in this study by the route Ri.
- t: stands for the system states set that are considered in this model by the under-study time shift periods during the day.
- a: stands for the decisions set that taken by the TS. In this work, the decisions of the TS are to nominate the best Ri that can cover the passengers requirements/expectations. Only six possible Rs between the proposed TSs are considered in our use case.
- r: stands for the immediate reward that the passenger will get according to taking decision a.

A $Q_i(t,a)$ is an action-state matrix that is formed to store the gathered reward/sanction for each state and the action pair of a route Ri at a time (t). Giving an example to that, this matrix reflects the predicted reward when takes an action (a) after using the route Ri at a time (t). The updating function of $Q_i(t,a)$ can be defined by:

$$Q_i(t,a) = (1-\alpha) * Q_i(t,a) + (1-\alpha) * (r_i(t) + \gamma * max_a{}' Q_i(t+1,a')) \quad (1)$$

Where:

- α: indicates the learning rate representing how fast the Q-values vary corresponding to the dynamic users reaction.
- γ: indicates the discount factor. It reflects if the immediate reward status against the future one. For the comparison, high value future reward is more valued than the immediate one.
- r_i: indicates the predicted immediate reward when choosing Ri at a time (t).
- $max_a{}' Q_i(t+1,a')$: represents the maximum predicted future reward once the system reaches a state $(t+1)$ when taking any decision action a.

Furthermore, once a passenger arrives the desired destination via Ri after the scheduled arrival-time, TS will get a sanction reaction. While, in case all the Rs are delayed, bonus is counted to the Rs of the minimum counted delay time.

$$r_i(t) = \frac{mt}{B_1(t)} + (\frac{mc}{B_2}) + B_3(t) + B_4(t) + (B_5(t) * Comfortability) + B_6(t) + B_7 \quad (2)$$

Where:

- mt is the reference trip duration of all Ri at a time (t) that represents the minimum travel time of Ri.
- B_1 parameter indicates the travel time due to catching Ri.
- B_2 parameter indicates the ticket price of the Ri comparative to a mean value (mc) that equals to one. It is assumed that, the ticket price plan is fixed during the week-days.
- B_3 and B_4 are bonus values for the Ri of minimum waiting and connection times respectively. These parameters are comparative between all the transport means and takes a value of 1 to the lowermost times up to 0.5 to the uppermost times.
- B_5 is the passengers satisfaction reaction in percentage reflecting the comfortability of using Ri at time (t). The default value equals to 1 as shown in Figure 3.

- B_6 is an added bonus to the Rs of low delays comparatively between them and on the contrary Ri of a highest delay has no bonus.

- B_7 is an indication to the green transport means which is preferred to be used than others as it cause less pollution. This parameter is a bonus totaled to the Ri of low transport means CO_2 emission.

Passengers reactions/feedbacks about the transport means are collected in real-time as reports. Each passenger is supposed to send his reaction/feedback using a dedicated simple mobile application that will supply the Decision System Algorithm $DSA(t)$ shown in Algorithm 1.

Algorithm 1 Q-Learning Algorithm: DSA(t)

Input: For each Ri, collect the passengers' feedback at their arrival destination
Input: $L1, \cdots,$ until $L6$ are 6 lists initialized by NULL
Input: Qi is initialized based on the previous knowledge of the traffic jam for every day and is rewarded based on the current time (t) by time interval (τ)
Input: Ni is the available Rs at the current time (t)
Output: The best route at time $(t+\tau)$
if $Ni \neq \emptyset$ **then**
  For$\{j,k,l,m,n,f\} \in Ni$ **do**
  {at time (t) there is a comparison between the available Rs as follows:}
  if $|RouteAvailable(j)| \neq 0$ and reward(j) > reward(i) for $(j \neq i)$ **then**
    $L1 \leftarrow j$
  end if
  if $|RouteAvailable(k)| \neq 0$ and reward(j) > reward(k) > reward(i) for $(k \neq i)$ and $(j \neq i)$ **then**
    $L2 \leftarrow k$
  end if
  if $|RouteAvailable(l)| \neq 0$ and reward(j) > reward(k) > reward(l) for $(k \neq i)$ and $(l \neq i)$ **then**
    $L3 \leftarrow l$
  end if
  if $|RouteAvailable(m)| \neq 0$ and reward(l) > reward(m) > reward(i) for $(l \neq i)$ and $(m \neq i)$ **then**
    $L4 \leftarrow m$
  end if
  if $|RouteAvailable(n)| \neq 0$ and reward(m) > reward(n) > reward(i) for $(m \neq i)$ and $(n \neq i)$ **then**
    $L5 \leftarrow n$
  end if
  if $|RouteAvailable(f)| \neq 0$ and reward(f) < reward(i) for $(k \neq i)$ for $f = j,k,l,m,n$ and $(f \neq i)$ **then**
    $L6 \leftarrow f$
  end if
  End for
  if $L1 \neq 0$ **then**
    The next Route is chosen from $L1$ of the largest Q-value
  else if $L1 = 0$ and $L2 \neq 0$ **then**
    The next Route is chosen from $L2$ of the largest Q-value
  else if $L1 = L2 = 0$ and $L3 \neq 0$ **then**
    The next Route is chosen from $L3$ of the largest Q-value
  else if $L1 = L2 = L3 = 0$ and $L4 \neq 0$ **then**
    The next Route is chosen from $L4$ of the largest Q-value
  else if $L1 = L2 = L3 = L4 = 0$ and $L5 \neq 0$ **then**
    The next Route is chosen from $L5$ of the largest Q-value
  else if $L1 = L2 = L3 = L4 = L5 = 0$ and $L6 \neq 0$ **then**
    The next Route is chosen from $L6$ of the largest Q-value
  else
    Ni is empty
    At time (t) a negative reward is generated for that Ri
  end if
end if
At time $(t+1)$ the reward for every Ri is computed after the interval (τ) based on step (2)
At time (t) the Q-value is updated based on Eq. (1)

3.3 Support Vector Machine Classification

SVM, a supervised machine learning tool, is used for classifying linear and non-linear dynamic systems. SVM decisions can be considered for the classification problems and the regression aspect. The best classification results from finding the best hyperplane fits data separation with error free and maximal distance between the closest vector to that hyperplane.

In this work, SVM is used to distinguish among two classes. Class one refers to the best route(s) to certain destination while the other class refers to the alternative routes that are available for use to the same destination.

The aim is to find a data separator between the two mentioned classes whatever linear or non-linear. From theoretical point of view, linear SVM matches our scenario as it is a question of only two classes. Assuming the routes data set (i.e. the collected traces) is D as:

$$D = \{(x^1, y^1), \cdots, (x^n, y^n)\}, x \in R^n, y \in (0,1) \quad (3)$$

Where:

- x is the set of training vectors.

- n is the number of days quarters under study that is equal to 25 period.

- y is the classs labels.

And the hyperplane is:

$$\langle w, x \rangle + b = 0 \quad (4)$$

Considering a canonical hyperplane [4], where the parameters w, b are constrained by,

$$min_i |\langle w, x^i \rangle + b| = 1 \quad (5)$$

The following constraints must be considered for a separating the canonical hyperplane, which is the most fitting assumption for non-linear classification,

$$y^i [\langle w, x^i \rangle + b] \geq 1, i = 1, \cdots, l \quad (6)$$

The distance $d(w,b;x)$ of a point x from the hyperplane (w,b) is:

$$d(w,b;x) = \frac{|\langle w, x^i \rangle + b|}{||w||} \quad (7)$$

For optimal hyperplane, the margin, ρ has to be maximized based on the constraints of (6) and given by:

$$\rho(w,b) = min_{x^i, y^i = -1} d(w,b;x^i) + min_{x^i, y^i = 1} d(w,b;x^i)$$

$$= \frac{2}{||w||} \quad (8)$$

So, the optimal hyperplane, the data separator that minimizes is:

$$\phi(w) = \frac{1}{2}||w||^2 \quad (9)$$

The optimization problem solution of (9) using the constraints of (6) is given by the saddle point of the Lagrange function:

$$\phi(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l}\alpha_i(y^i[\langle w,x^i\rangle + b] - 1) \quad (10)$$

where α is the Lagrange multiplier that must be minimized with respect to w, b and maximized with respect to $\alpha \geq 0$. This dual problem is given by:

$$max_\alpha W(\alpha) = max_\alpha(min_{w,b}\phi(w,b,\alpha)) \quad (11)$$

To get the minimum w and b of the Lagrangian, ϕ is given by,

$$\frac{\partial\phi}{\partial b} = 0 \Rightarrow \sum_{i=1}^{l}\alpha_i y_i = 0$$

$$\frac{\partial\phi}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{l}\alpha_i y_i x_i \quad (12)$$

So, from (10), (11), and (12), the dual problem expressed by:

$$max_\alpha W(\alpha) = max_\alpha - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j$$

$$\langle x_i, x_j\rangle + \sum_{k=1}^{l}\alpha_k \quad (13)$$

and the solution of this problem will be given as follows:

$$\alpha^* = arg\ min_\alpha \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j\langle x_i, x_j\rangle - \sum_{k=1}^{l}\alpha_k \quad (14)$$

where $\alpha_i \geq 0$, for $i = 1,\cdots,l$.

By solving (14) according to its constraints defines the Lagrange multipliers then, obtains the optimal separating hyperplane as well as follows:

$$w^* = \sum_{i=1}^{l}\alpha_i y_i x_i$$

$$b^* = -\frac{1}{2}\langle w^*, x_r + x_s\rangle \quad (15)$$

where x_r and x_s are the support vectors from each class satisfying:

$$\alpha_r, \alpha_s > 0, \quad y_r = 0, y_s = 1 \quad (16)$$

and the hard classifier is:

$$f(x) = sgn(\langle w^*, x\rangle + b) \quad (17)$$

The input vectors classification that refer to the routes parameters can be solved as a nonlinear problem that fits our case as well. This depends on the input vectors distribution. For a general form in high dimensional feature space, lets suppose SVM maps the input vector x into a high dimensional feature space, z. So, the optimization problem becomes:

$$\alpha^* = arg\ min_\alpha \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{k=1}^{l}\alpha_k \quad (18)$$

where K is the kernel function for non-linear mapping into feature space with the same limitations, which are:

$$0 \leq \alpha_i \leq C, i = 1,\cdots,l\sum_{j=1}^{l}\alpha_j y_j = 0 \quad (19)$$

where C is a regularization parameter of the kernel function and represents the upper bound on the permissible values. Solving (18) with its constraints in (19), resulting the Lagrange multipliers, and a hard classifier as follows:

$$f(x) = sgn(\sum_{i\in SV_s}\alpha_i y_i K(x_i, x) + b) \quad (20)$$

where:

$$\langle w^*, x\rangle = \sum_{i=1}^{l}\alpha_i y_i K(x_i, x)$$

$$b^* = -\frac{1}{2}\sum_{i=1}^{l}\alpha_i y_i[K(x_i, x) + K(x_i, x_r)] \quad (21)$$

4 Performance Validation

This section demonstrates the proposed GTDS model performance evaluations. It is evaluated based on using Q-Learning and SVM learning methods. *Matlab* 2012 tool is used for the simulation part with the following conditions:

- The model use case under study considers Paris region in France. The departure station is (*Gare De Paris Montparnasse*) and the arrival destination station is the (*Gare De Paris Nord*) as the most two famous stations in this Parisian region.
- The routes *Rs* available between the source and destination stations are many but, we processed and filtered them to form six routes only of closed results for simple comparisons.
- Time shift under study is divided into two sets. Set 1 is representing the two days of the weekend (i.e. Saturday and Sunday) and set 2 is representing the two selected working days which are Monday and Friday.
- Time shift under study starts at 7:00 AM till 13:00 PM
- The ticket price of *Ri* is supposed to be unity and its value is redeemed comparatively based on the transport mean used. The route cost is supposed to be the same as this fits the case in France that depends on dividing the destinations to zones.

- The collected traces of travel duration, waiting and connection times are gathered from the real-time schedules of the *SNCF Transilien* official site [5].

- The passengers comfortability reactions/feedbacks of the transport means understudy follows the chart in Figure 3. This comprises many parameters such as the capacity, the design satisfaction, the rate of flow (periodicity), and the ease of catching each of them.

- The actual CO_2 emission is measured in gram per person of each *Ri* and gathered from *Vianavigo* site [6].

- Number of stops: It is dynamic parameter. Its value depends on the route combinations that may comprise a combination of different transport means.

The simulations are categorized in two parts; part one is concerned about the SVM learning algorithm, and part two focuses on the Q-learning algorithm. In part one, the used traces of the determined source and destination are gathered for two weeks just for a proof of concept from [5]. We considered the first week traces as the training set and the second week traces as the testing set.

After the validation of SVM, the obtained testing set results are shown in Figure 4 and Figure 5. They illustrate the SVM accuracy calculated for both the week-end and working days respectively.

As illustrated in these figures (Figure 4 and Figure 5), the SVM classification accuracy is looking good in week-end days but has less accuracy in the working days. These are anticipated results as working days may have rush hours that influence on choosing the best route. Despite the results disorder, the system performs steady as the error is so small that is closed to one over six.

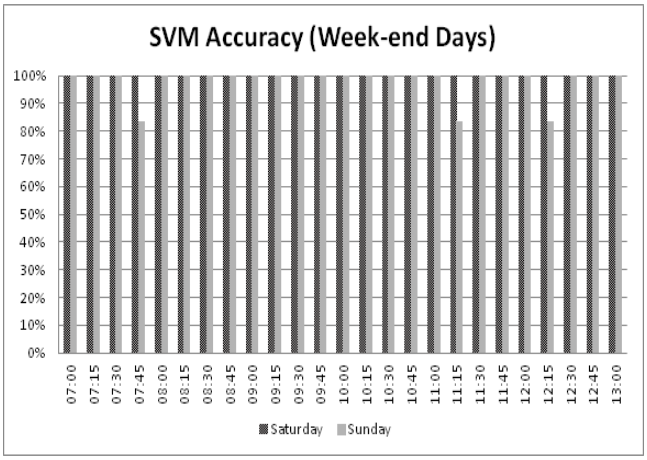


Figure 4: The output SVM system testing accuracy during week-end tested days, for the X-axis is represented the time shift from 7:00 AM to 13:00 PM as testing period

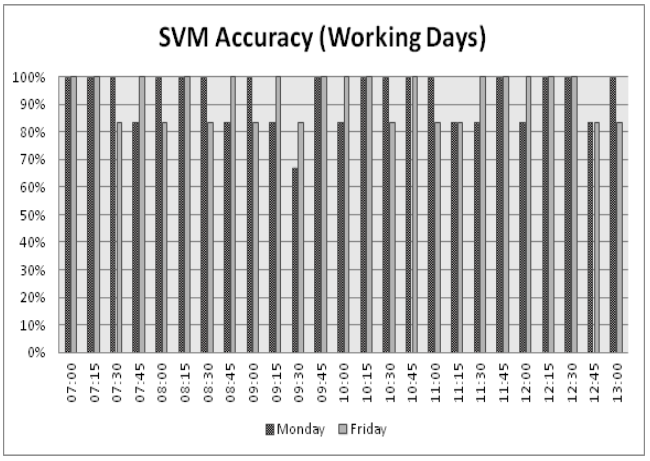


Figure 5: The output SVM system testing accuracy during working tested days, for the X-axis is represented the time shift from 7:00 AM to 13:00 PM as testing period

The second simulation part is concerned to the Q-learning machine results of choosing the best route *Ri*. The same traces and system settings are used for the Q-learning simulation. This simulation considers two use cases as follows:

A) The week-end days: it concerns to the delay of using each transport mean on one of the week-end days (Saturday) as shown in Figure 6. The use of bus compared to Metro and Train has the largest delay at the same day quarter although roads are not easily suffering of congestion during the week-ends. Moreover, the metros are not far from buses statistics due to their limited circulations in the weekend days.

B) Working-days: this simulation concerns to one working day like Friday traffic situation. The highest rush hours of the day are simulated. Figure 7 shows each transport mean delay. As clear in the figure, the use of the bus causes the largest delay with respect to Metro and Train at the same day quarter due to the traffic jams on normal routes.

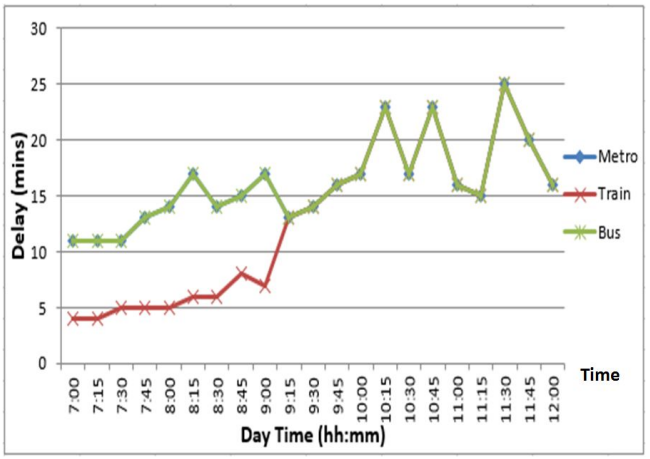


Figure 6: Saturday travel delay due to using either Metro, Train, or Bus. The hh: is the hour while mm: represents the minutes part

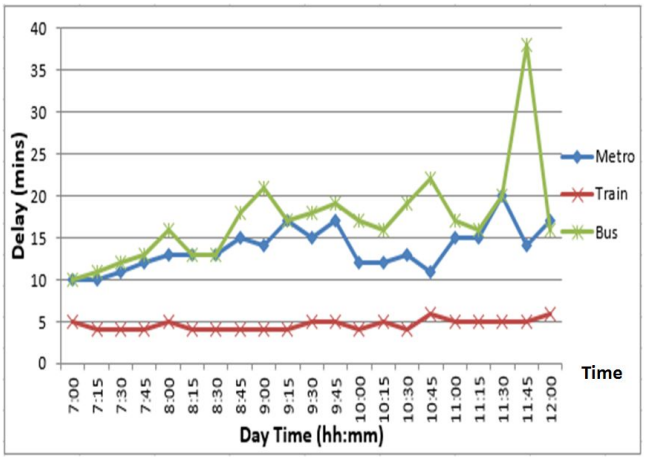


Figure 7: Friday travel delay due to using either Metro, Train, or Bus. The hh: is the hour while mm: represents the minutes part

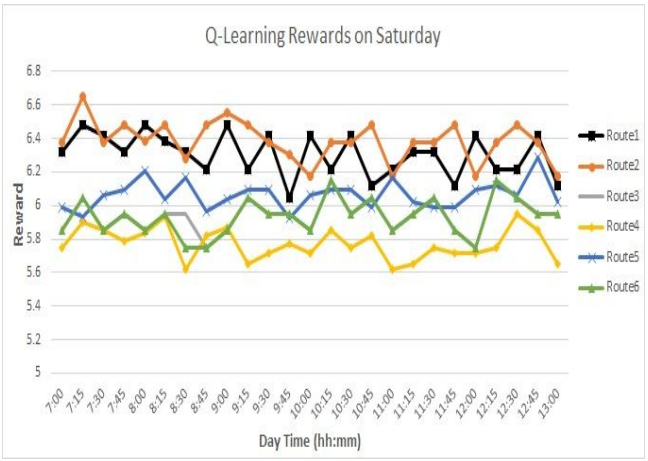


Figure 8: The rewards for Q-Learning algorithm based Saturday testing set. The hh: is the hour while mm: represents the minutes part

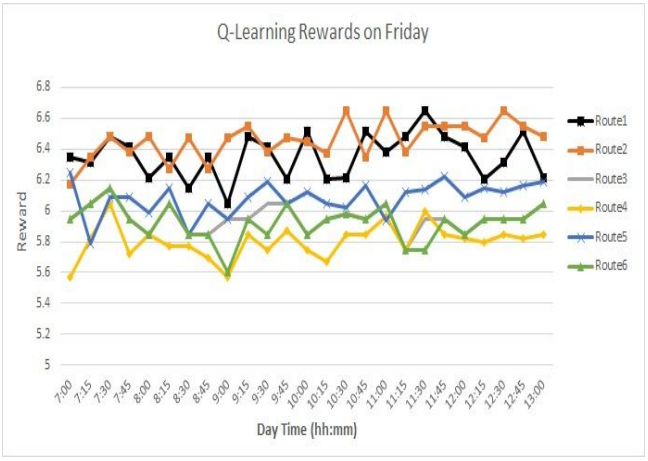


Figure 9: The rewards for Q-Learning algorithm based Friday testing set. The hh: is the hour while mm: represents the minutes part

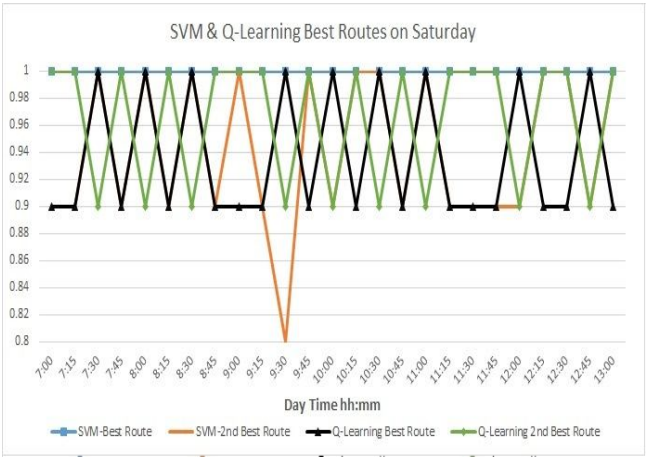


Figure 10: SVM versus Q-Learning algorithms results for the week-end selected day as Saturday

The results shown in Figure 8 and Figure 9 demonstrate the Q-values which stands for the rewards of the optional routes of each day quarter on Saturday and Friday respectively. As clear in the figures, the best route decision changes with respect to the day quarter time and routes conditions. These routes are composed of different two or more traffic means. Therefore, these figures illustrate that, the recommended routes to be used along the studied day hours depends on the conditions and delays for each route which is composed of different connections between the three-transport means (metro, bus and train).

Figure 10 and Figure 11 show the best two routes suggested by the two proposed machine learning systems (i.e. SVM and Q-Learning) for Saturday and Friday respectively. As appear in these figures, each learning algorithm decision is various due to the difference of the learning technique of both algorithms. Also, due to the difference of the input parameters weight.And finally, because SVM is a supervised learning model but, Q-learning is based on reinforcement learning.

The main purpose from these two figures is to clarify the criteria of each algorithm output and shows that each algorithm may fits according to certain applications. If the ITS system is merely in stable with no fluctuates, SVM would be a perfect fit for these scenarios. In contrast, Q-learning could suitable for infrequent cases such as accidents. This is because of Q-learning depends on conditions of the current collected data in addition to its decision history of the previous day while SVM depends on the training sets.

5 Conclusions

The framework of route selection in the ITS multimodal stations is investigated in this paper using context awareness decision model based on machine learning tools. Through this work, we used two different models; the Q-Learning as a reinforcement model and the SVM as a supervised model.

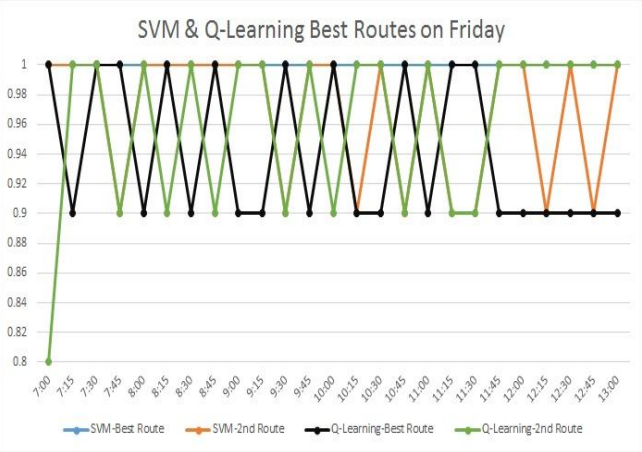


Figure 11: SVM versus Q-Learning algorithms results for the selected work-day as Friday

In the first phase, we built the Q-Learning to select among six studied routes the best one. After that, we used SVM model in order to first classify the proposed routes, second to calculate the model's accuracy, and third to predict the best solution for the upcoming routes. The obtained results from the simulation part of both approaches indicated the advantages and disadvantages for each technique and when (time during the day or day during the week) or where (according traffic status) we can use them in ITS stations.

Actually, Q-Learning approach is best choice that can be used when the ITS system is not stable and subject to many disturbances while, SVM model is better used when the ITS system is stable and the variations over the time in data sets are steady. That is why we recommend using hybrid solutions to face such cases when searching the optimal classification based on different time intervals and with green parameters. On the positive side, ITS will enable the Green Transport Decision System (GTDS) to decide based on context-sensitive information the best route, to improve the efficiency of multimodal stations usage and to reduce the environmental impacts by including the CO_2 effects.

In the future work, it will be more realistic to focus on three dimensional model approach that will consider in addition to user-location; the time. This spatio-temporal model can be applied in different universal applications. But, the problem is the difficulty of having such kind of statistics as it is subject to country regulations related to security and privacy issues. However, it can provide a new generation of traffic flows modeling that will add dynamic, predicative and adaptive control for multimodal traffic.

Acknowledgment

This manuscript is carried out under the Franco-Egyptian research program called IMHOTEP Project (call: 2016-2017), as a scientific collaboration between Campus France (represented by Telecom SudParis) and the Academy of Scientific Research and Technology (ASRT) Egypt.[1]

References

[1] A. M. Said, E. Abd-Elrahman and H. Afifi, "A comparative study on machine learning algorithms for green context-aware intelligent transportation systems," 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, 2017, pp. 1-5.

[2] A. M. Said, A. Soua, E. Abd-Elrahman, and H. Afifi, "Context-Aware Multi-modal Traffic Management in ITS: A Q-Learning based Algorithm," International Wireless Communications and Mobile Computing Conference (IWCMC 2015), Dubrovnik-Croatia, Aug. 2015, pp. 674-679.

[3] V. N. Vapnik, "The Nature of Statistical Learning Theory," New York: Springer, 1995.

[4] S. R. Gunn, "Support Vector Machines for Classification and Regression," Technical report:Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 1998.

[5] SNCF Transilien 2018. [Online]. Available: https://www.transilien.com/en. [Accessed: 29-Sep-2018].

[6] Vianavigo 2018. [Online]. Available: https://www.vianavigo.com/en/home. [Accessed: 29-Sep-2018].

[7] S.J. Russel and P. Norvig, "Artificial Intelligence: A Modern Approach," Third Edition, Prentice Hall, 2010.

[8] D. Fagan and R. Meier, "Intelligent time of arrival estimation," IEEE Forum on Integrated and Sustainable Transportation System (FISTS), Vienna, 2011, pp.60 - 66.

[9] S. El-Tantawy and B. Abdulhai, "Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC)," 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), Anchorage, 2012, pp. 319-326.

[10] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC): Methodology and Large-Scale Application on Downtown Toronto," IEEE Transactions on Intelligent Transportation Systems, Vol.14, No. 3, 2013, pp. 1140-1150.

[11] S. El-Tantawy, and and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control," 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), Funchal, 2010, pp. 665-670.

[12] I. Dusparic and V. Cahill, "Distributed W-Learning: Multi-Policy Optimization in Self-Organizing Systems," Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems SASO '09, San Francisco, 2009, pp. 20-29.

[13] J. Song and Z. Jin, "Q-learning based multi-intersection traffic signal control model," International Conference on System Science, Engineering Design and Manufacturing Informatization (ICSEM), Vol.2, 2011, pp. 280-283.

[14] J.Song, Z. Jin, and W. Zhu, "Implementing traffic signal optimal control by multiagent reinforcement learning," International Conference on Computer Science and Network Technology (ICCSNT), Vol. 4, Harbin, 2011, pp.2578-2582.

[15] A. Karatzoglou, D. Meyer, and K. Hornik, "Support Vector Machines in R," Journal of Statistical Software, Volume 15, Issue 9, 2006, pp.1-28.

[16] C. Wu, J. Ho, and D. T. Lee, "Travel-Time Prediction With Support Vector Regression," IEEE Transactions on Intelligent Transportation Systems, Vol. 5, No. 4, 2004, pp. 276-281.

[17] C. Wu, C. Wei, D. Su, M. Chang, and J. Ho, "Travel Time Prediction with Support Vector Regression," IEEE Intelligent Transportation Systems, Vol.2, 2003, pp. 1438-1442.

[1] Programme Imhotep (2016-2017): http://www.campusfrance.org/fr/imhotep

[18] A. Yusuf, and V. Madisetti, "Configuration for Predicting Travel-Time Using Wavelet Packets and Support Vector Regression," Journal of Transportation Technologies, Vol.3, No. 3, 2013, pp. 220-231.

[19] Y. Chen, H. J. van Zuylen, and Y. Qipeng, "Travel Time Prediction on Urban Networks Based on Combining Rough Set with Support Vector Machine," International Conference on Logistics Systems and Intelligent Management, Vol. 1, 2010, pp. 586-589.

[20] S. Zhong, J. Hu, S. Ke, X. Wang, J. Zhao, and B. Yao, "A Hybrid Model based on Support Vector Machine for Bus Travel-Time Prediction," PROMET Traffic and Transportation Journal, Vol. 27, No. 4, 2015, pp. 291-300.

[21] C. Bai, Z. Peng, Q. Lu, and J. Sun, "Dynamic Bus Travel Time Prediction Models on Road with Multiple Bus Routes," Computational Intelligence and Neuroscience (Hindawi Publishing Corporation), Vol.2015, No.432389, 2015.

[22] H. Zhou, Y. Qin, and Y. Li, "A Partial Least Square Based Support Vector Regression Rail Transit Passenger Flow Prediction Method," International Journal of u- and e- Service, Science and Technology, Vol.7, No.2, 2014, pp. 101-112.

[23] S. Maldonado-Bascn, S. Lafuente-Arroyo, P. Gil-Jimnez, H. Gmez-Moreno, and F. Lpez-Ferreras, "Road-Sign Detection and Recognition Based on Support Vector Machines," IEEE Transactions on Intelligent Transportation Systems, Vol.8, No.2, 2007, pp. 264-278.

[24] S. Sathiya, M. Balasubramanian, and S. Palanivel, "Pattern Recognition Based Detection Recognition of Traffic Sign Using SVM," International Journal of Engineering and Technology (IJET), Vol. 6, No. 2, 2014, pp. 1147-1157.

[25] S. B.Wali, M. A. Hannan, A. Hussain, and S. A. Samad, "An Automatic Traffic Sign Detection and Recognition System Based on Colour Segmentation, Shape Matching, and SVM," Mathematical Problems in Engineering (Hindawi Publishing Corporation), Vol.2015, No.250461, 2015.

[26] S. Lafuente-Arroyo et al., "A decision support system for the automatic management of keep-clear signs based on support vector machines and geographic information systems," Expert Systems with Applications, Vol.37, 2010, pp. 767-773.

[27] D. Kim, "Prediction Performance of Support Vector Machines with Fused Data in Road Scene Analysis," International Journal of Transportation, Vol.3, No.3, 2015, pp.41-48.

[28] C.C Lin, C.W Lin, D. Huang, and Y. Chen, "Design a Support Vector Machine-based Intelligent System for Vehicle Driving Safety Warning," 11th International IEEE Conference on Intelligent Transportation Systems, 2008, pp. 938-943.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 2, 329-345 (2018)
www.astesj.com
Special Issue on Advancement in Engineering Technology

ASTES Journal
ISSN: 2415-6698

Efficient Alignment of Very Long Sequences

Chunchun Zhao*, Sartaj Sahni

Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA

A R T I C L E I N F O

Article history:
Received: 20 March, 2018
Accepted: 16 April, 2018
Online: 30 April, 2018

Keywords:
Sequence alignment
Energy efficient
Multi core

A B S T R A C T

We consider the problem of aligning two very long biological sequences. The score for the best alignment may be found using the Smith-Waterman scoring algorithm while the best alignment itself may be determined using Myers and Miller's alignment algorithm. Neither of these algorithms takes advantage of computer caches to obtain high efficiency. We propose cache-efficient algorithms to determine the score of the best alignment as well as the best alignment itself. All algorithms were implemented using C and OpenMP, and benchmarked using real data sets from the National Center for Biotechnology Information (NCBI) database. The test computational platforms were Xeon E5 2603, I7-x980 and Xeon E5 2695. Our best single-core cache-efficient scoring algorithm reduces the running time by as much as 19.7% relative to the Smith-Waterman scoring algorithm and our best cache-efficient alignment algorithm reduces the running time by as much as 17.1% relative to the Myers and Miller alignment algorithm. Multicore versions of our cache-efficient algorithms scale quite well up to the 24 cores we tested; achieving a speedup of 22 with 24 cores. Our multi-core scoring and alignment algorithms reduce the running time by as much as 61.4% and 47.3% relative to multi-core versions of the Smith-Waterman scoring algorithm and Myers and Miller's alignment algorithm, respectively.

1 Introduction

Sequence alignment is a fundamental and well-studied problem in the biological sciences. In this problem, we are given two sequences $A[1:m] = a_1 a_2 \cdots a_m$ and $B[1:n] = b_1 b_2 \cdots b_n$ and we are required to find the score of the best alignment and possibly also an alignment with this best score. When aligning two sequences, we may insert gaps into the sequences. The score of an alignment is determined using a matching (or scoring) matrix that assigns a score to each pair of characters from the alphabet in use as well as a gap penalty model that determines the penalty associated with a gap sequence. In the linear gap penalty model, the penalty for a gap sequence of length $k > 0$ is kg, where g is some constant while in the affine model this penalty is $g_{open} + (k-1) * g_{ext}$. The affine model more accurately reflects the fact that opening a gap is more expensive than extending one. Two versions of sequence alignment–global and local–are of interest. In global alignment, the entire A sequence is to be aligned with the entire B sequence while in local alignment, we wish to find a substring of A and B that have the highest alignment score. The alphabet for DNA, RNA, and protein sequences is, respectively, {A, T, G, C}, {A, U, G, C}, and {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y}.

Figure 1 illustrates these concepts using the DNA sequences $A[1:8] = \{\text{AGTACGCA}\}$ and $B[1:5] = \{\text{TATGC}\}$. The symbol '_' denotes the gap character. The alignment of Figure 1(a) is a global alignment and that of Figure 1(b) is a local one. To score the alignments, we have used the linear penalty model with $g = -2$ and the scores for pairs of aligned characters, which are taken from BLOSUM62 matrix in [1], are $c(T,T) = 5$, $c(A,A) = 4$, $c(C,C) = 9$, $c(G,G) = 6$, and $c(C,T) = -1$. The score for the shown global alignment is 17 while that for the shown local alignment is 23. If we were using an affine penalty model with $g_{open} = -4$ and $g_{ext} = -2$, then the penalty for each of the gaps in positions 1 and 8 of the global alignment would be -4 and the overall score for the global alignment would be

*Corresponding Author: Chunchun Zhao, czhao@cise.ufl.edu

https://dx.doi.org/10.25046/aj030236

13.

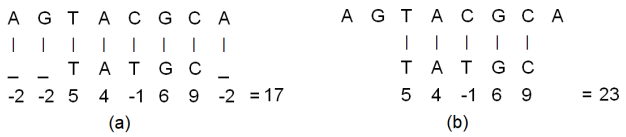


Figure 1: Example alignments using the linear gap penalty model. (a) Global alignment (b) Local alignment

In [2], the authors first proposed an $O(mn)$ time algorithm, called Needleman-Wunsch(NW) algorithm, for global alignment using the linear gap model. This algorithm requires $O(n)$ space when only the score of the best alignment is to be determined and $O(mn)$ space when the best alignment is also to be determined. In [3], the authors proposed a new algorithm called Smith-Waterman(SW) algorithm, which modified the NW algorithm so as to determine the best local alignment. In [4], the author proposed a dynamic programming algorithm called Gotoh algorithm, for sequence alignment using an affine gap penalty model. The asymptotic complexity of the SW and Gotoh algorithms is the same as that of the NW algorithm.

When mn is large and the best alignment is sought, the space, $O(mn)$, required by the algorithms of NW, SW and Gotoh exceeds what is available on most computers. The best alignment for these large instances can be found in [5] using sequence alignment algorithms derived from Hirschberg's linear space divide-and-conquer algorithm for the longest common subsequence problem. In [6], the authors developed a Myers-Miller alignment algorithm. It is the linear space $O(mn)$ time version of Hirschberg's algorithm for global sequence alignment using an affine gap penalty model. And in [7], the authors do this for local alignment.

In an effort to speed sequence alignment, fast sequence-alignment heuristics have been developed. As in [8, 9, 10],BLAST, FASTA, and Sim2 are a few examples of software systems that employ sequence alignment heuristics. Another direction of research, also aimed at speeding sequence alignment, has been the development of parallel algorithms. Parallel algorithms for sequence alignment may be found in [11]-[19], for example.

In this paper, we focus on reducing the number of cache misses that occur in the computation of the score of the best alignment as well as in determining the best alignment. Although we explicitly consider only the linear gap penalty model, our methods readily extend to the affine gap penalty model. Our interest in cache misses stems from two observations–(1) the time required to service a last-level-cache (LLC) miss is typically 2 to 3 orders of magnitude more than the time for an arithmetic operation and (2) the energy required to fetch data from main memory is typically between 60 to 600 times that needed when the data is on the chip. As a result of observation (1), cache misses dominate the overall running time of applications for which the hardware/software cache prefetch modules on the target computer are ineffective in predicting future cache misses. The effectiveness of hardware/software cache prefetch mechanisms varies with application, computer, and compiler. So, if we are writing code that is to be used on a variety of computer platforms, it is desirable to write cache-efficient code rather than to rely exclusively on the cache prefetching of the target platform. Even when the hardware/software prefetch mechanism of the target platform is very effective in hiding memory latency, observation (2) implies excessive energy use when there are many cache misses.

This paper is an extension of work originally in [20], which has been presented by us in the 2015 IEEE 5th international conference on Computational Advances in Bio and Medical Sciences (ICCABS). The main contributions are

1. cache efficient single-core and multi-core algorithms to determine the score of the best alignment;

2. cache efficient single-core and multi-core algorithms to determine the best alignment.

The rest of the paper is organized in the following way. In Section 2, we describe our cache model. Our cache-efficient algorithms for scoring and alignment are developed and analyzed in Section 3. Experimental results are presented in Section 4. In Section 5, we present a discussion of these results and in Section 6, we present the limitations of our work. Finally, we conclude in Section 7.

2 Cache Model

For simplicity in the analysis, we assume a single cache comprised of s lines of size w words (a word is large enough to hold a piece of data, typically 4 bytes) each. So, the total cache capacity is sw words. The main memory is partitioned into blocks also of size w words each. When the program needs to read a word that is not in the cache, a cache miss occurs. To service this cache miss, the block of main memory that includes the needed word is fetched and stored in a cache line, which is selected using the LRU (least recently used) rule. Until this block of main memory is evicted from this cache line, its words may be read without additional cache misses. We assume the cache is written back with write allocate. Write allocate means that when the program needs to write a word of data, a write miss occurs if the block corresponding to the main memory is not currently in cache. To service the write miss, the corresponding block of main memory is fetched and stored in a cache line. Write back means that the word is written to the appropriate cache line only. A cache line with changed content is written back to the main memory when it is about to be overwritten by a new block from main memory.

Rather than directly assess the number of read and write misses incurred by an algorithm, we shall count the number of read and write accesses to main memory.

Every read and write miss makes a read access. A read and write miss also makes a write access when the data in the replacement cache line is written to main memory.

We emphasize that the described cache model is a very simplified model. In practice, modern computers commonly have two or three levels of cache and employ sophisticated adaptive cache replacement strategies rather than the LRU strategy described above. Further, hardware and software cache prefetch mechanisms are often deployed to hide the latency involved in servicing a cache miss. These mechanisms may, for example, attempt to learn the memory access pattern of the current application and then predict the future need for blocks of main memory. The predicted blocks are brought into cache before the program actually tries to read/write from/into those blocks thereby avoiding (or reducing) the delay involved in servicing a cache miss. Actual performance is also influenced by the compiler used and the compiler options in effect at the time of compilation. As a result, actual performance may bear little relationship to the analytical results obtained for our simple cache model. Despite this, we believe the simple cache model serves a useful purpose in directing the quest for cache-efficient algorithms that eventually need to be validated experimentally.

3 Cache Efficient Algorithms

3.1 Scoring Algorithms

3.1.1 Needleman-Wunsch and Smith-Waterman algorithm

Let H_{ij} be the score of the best global alignment for $A[1:i]$ and $B[1:j]$. We wish to determine H_{mn}. In [2], the authors derived the following dynamic programming equations for H using the linear gap penalty model. These equations may be used to compute H_{mn}.

$$H_{i,0} = -i * g \; H_{0,j} = -j * g, \; 0 \le i \le m, \; 0 \le j \le n \quad (1)$$

When $i > 0$ and $j > 0$,

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + c(a_i, b_j) \\ H_{i,j-1} + c(_, b_j) = H_{i,j-1} - g \\ H_{i-1,j} + c(a_i, _) = H_{i-1,j} - g \end{cases} \quad (2)$$

where $c(a_i, b_j)$ is the match score between characters a_i and b_j and g is the gap penalty.

For local alignment, H_{ij} denotes the score of the best local alignment for $A[1:i]$ and $B[1:j]$. In [3], the Smith-Waterman equations for local alignment using the linear gap penalty model are:

$$H_{i,0} = 0, \; H_{0,j} = 0, \; 0 \le i \le m, \; 0 \le j \le n \quad (3)$$

When $i > 0$ and $j > 0$,

$$H_{i,j} = \max \begin{cases} 0 \\ H_{i-1,j-1} + c(a_i, b_j) \\ H_{i,j-1} + c(_, b_j) = H_{i,j-1} - g \\ H_{i-1,j} + c(a_i, _) = H_{i-1,j} - g \end{cases} \quad (4)$$

Several authors (in [5, 6], for example) have observed that the score of the best local alignment may be determined using a single array $H[0:n]$ as in algorithm *Score*(Algorithm 1.)

Algorithm 1 Smith-Waterman scoring algorithm

```
1:  Score(A[1:m], B[1:n])
2:  for j ← 0 to n do
3:     H[j] ← 0 //Initialize row 0
4:  end for
5:  for i ← 1 to m do
6:     diag ← 0 // Compute row i
7:     for j ← 1 to n do
8:        nextdiag ← H[j]
9:        H[j] ← max{0, diag + c(A[i], B[j]), H[j − 1] −
          g, H[j] − g}
10:       diag ← nextdiag
11:    end for
12: end for
13: return H[n]
```

The scoring algorithm for the Needleman and Wunsch algorithm is similar. It is easy to see that the time complexity of the algorithm of Algorithm 1 is $O(mn)$ and its space complexity is $O(n)$.

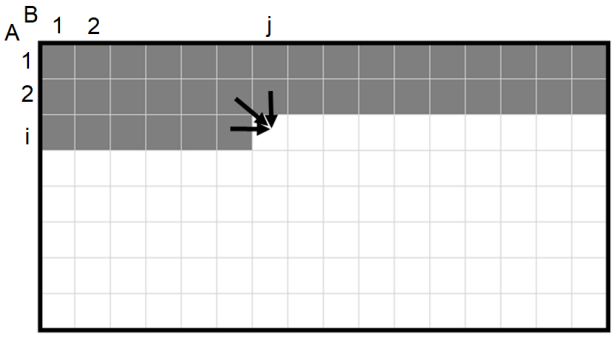


Figure 2: Memory access pattern for Score algorithm(Algorithm 1).

For the (data) cache miss analysis, we focus on read and write misses of the array H and ignore misses due to the reads of the sequences A and B as well as of the scoring matrix c (notice that there are no write misses for A, B, and c). Figure 2 shows the memory access pattern for H by algorithm *Score*. The first row denotes the initialization of H and subsequent rows denote access for different value of i (i.e., different iterations of the `for i` loop). The computation of H_{ij} is done using a single one-dimensional array $H[]$, following the i'th iteration of the `for i` loop, $H[j] = H_{ij}$. In each iteration of this loop, the elements of $H[]$ are accessed left-to-right. During the initialization loop, H is brought into the cache in blocks of size w. Assume that n is sufficiently large so that $H[]$ does not entirely fit into the cache. Hence, at some value of j, the cache capacity is reached and further progress of the initialization loop causes the least recently used blocks of $H[]$

(i.e., blocks from left to right) to be evicted from the cache. The evicted blocks are written to main memory as they have been updated. So, the initialization loop results in n/w read accesses and (approximately) n/w write accesses (the number of write accesses is actually $n/w - s$). Since the left part of $H[]$ has been evicted from the cache by the time we start the computation for row i ($i > 0$), each iteration of the `for i` loop also results in n/w read accesses and approximately n/w write accesses. So, the total number of read accesses is $(m+1)n/w \approx mn/w$ and the number of write accesses is also $\approx mn/w$. The number of read and write accesses is $\approx 2mn/w$, when n is large.

We note, however, that when n is sufficiently small that $H[]$ fits into the cache, the number of read accesses is n/w (all occur in the initialization loop) and there are no write accesses. In practice, especially in the case of local alignment involving a very long sequence, one of the two sequences A and B is small enough to fit in the cache while the other may not fit in the cache. So, in these cases, it is desirable to ensure that A is the longer sequence and B is the shorter one so that H fits in the cache entirely. This is accomplished by swap A and B sequences.

When $m < n$ and $H[1:m]$ fits into the cache and $H[1:n]$ does not, algorithm *Score* incurs $O(mn/w)$ read/write accesses, while swap A and B incurs $O(m/w)$ read/write accesses.

3.1.2 Diagonal Algorithm

An alternative to computing the score by rows is to compute by diagonals. While this uses two one-dimensional arrays rather than one, it is more readily parallelized than *Score* as all elements on a diagonal can be computed at the same time; elements on a row need to be computed in sequence.

Algorithm 2 Diagonal scoring algorithm

```
1:  Diagonal(A[1:m], B[1:n])
2:  d2[0] ← d1[0] ← d1[1] ← 0
3:  for d ← 2 to m + n do
4:      x ← (d <= m ? 0 : d − m); y ← (d <= n ? d : n);
5:      for i ← x to y do
6:          j ← d − i
7:          diag ← next + c(A[i] + B[j])
8:          left ← d1[i − 1] − g
9:          up ← d1[i] − g
10:         next ← d2[i]
11:         d2[i] ← max{0, diag, left, up}
12:     end for
13:     swap(d1, d2)
14: end for
15: return d1[n]
```

Algorithm *Diagonal* (Algorithm 2) uses two one-dimensional arrays $d1[]$ and $d2[]$, where $d2[]$ stores the scores for the $(d-2)$th diagonal data and $d1[]$ stores them for the $(d-1)$th diagonal. When we compute the element $H_{i,j}$ in the dth diagonal, the previous diagonal element $H_{i-1,j-1}$ is fetched from $d2[]$ and the previous left element $H_{i-1,j}$ and previous upper element $H_{i,j-1}$ are fetched from $d1[]$. The calculated $H_{i,j}$ overwrites the old value in $d2[]$.

The total number of read accesses is mn/w for each diagonal array and the total number of write accesses is mn/w for both arrays combined. The number of cache misses for *Diagonal* is approximately $3mn/w$ when n is large.

3.1.3 Strip Algorithm

When neither $H[1:m]$ nor $H[1:n]$ fits into the cache, accesses to main memory may be reduced by computing H_{ij} by strips of width q such that q consecutive elements of $H[]$ fit into the cache. Specifically, we partition $H[1:n]$ into n/q strips of size q (except possibly the last strip whose size may be smaller than q) as in Figure 3. First, all H_{ij} in strip 0 are computed, then those in strip 1, and so on. When computing the values in a strip, we need those in the rightmost column of the preceding strip. So, we save these rightmost values in a one-dimensional array $strip[0:m]$. The algorithm is given in Algorithm 3. We note that sequence alignment by strips has been considered before. For example, in [12], the authors using the similar approach in their GPU algorithm. Their use differs from ours in that they compute the strips in pipeline fashion with each strip assigned to a different pipeline stage in round robin fashion and within a strip, the computation is done by anti-diagonals in parallel. On the other hand, we do not pipeline the computation among strips and within a strip, our computation is by rows.

Algorithm 3 Strip scoring algorithm

```
1:  Strip(A[1:m], B[1:n])
2:  for j ← 1 to m do
3:      strip[j] ← 0 //leftmost strip
4:  end for
5:  for t ← 1 to n/q do
6:      for j ← t*q to t*q + q − 1 do
7:          H[j] ← 0 //Initialize first row
8:      end for
9:      for i ← 1 to m do
10:         diag ← strip[i − 1]
11:         H[t*q − 1] ← strip[i]
12:         for j ← t*q to t*q + q − 1 do
13:             nextdiag ← H[j]
14:             H[j] ← max{0, diag + c(A[i], B[j]), H[j − 1] −
                g, H[j] − g}
15:             diag ← nextdiag
16:         end for
17:         strip[i] ← H[t*q + q − 1]
18:     end for
19: end for
20: return H[n]
```

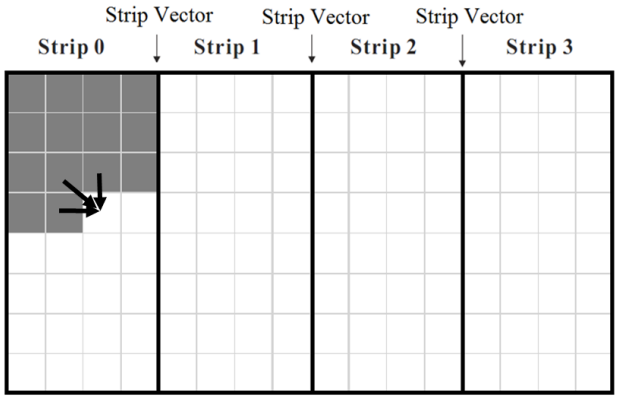


Figure 3: Memory access pattern for Strip algorithm (Algorithm 3).

It is easy to see that the time complexity of algorithm *Strip* is $O(mn)$ and that its space complexity is $O(m+n)$. For the cache misses, we focus on those resulting from the reads and writes of $H[]$ and $strip[]$. The initialization of *strip* results in m/w read accesses and approximately the same number of write accesses. The computation of each strip makes the following accesses to main memory:

1. q/w read accesses for the appropriate set of q entries of H for the current strip and q/w write accesses for the cache lines whose data are replaced by these H values. The write accesses are, however, not made for the first strip.

2. m/w read accesses for *strip* and m/w write accesses. The number of write accesses is less by s for the last strip.

So, the overall number of read accesses is $m/w + (q/w + m/w) * n/q = m/w + n/w + mn/(wq)$ and the number of write accesses is approximately the same as this. So, the total number of main memory accesses is $\approx 2mn/(wq)$ when m and n are large.

3.2 Alignment Algorithms

In this section, we examine algorithms that compute the alignment that results in the best score rather than just the best score. While in the previous section we explicitly considered local alignment and remarked that the results readily extend to global alignment, in this section we explicitly consider global alignment and remark that the methods extend to local alignment.

3.2.1 Myers and Miller's Algorithm

When aligning very long sequences, the $O(mn)$ space requirement of the full-matrix algorithm exceeds the available memory on most computers. For these instances, we need a more memory-efficient alignment algorithm. In [6], Myers and Miller have adapted Hirschberg's linear space algorithm for the longest common subsequence problem to find the best global alignment in linear space. Its time complexity is $O(mn)$. However, this linear space adaptation performs about twice as many operations as does the full-matrix algorithm. In [11], the authors have developed a hybrid algorithm, FastLSA, whose memory requirement adapts to the amount of memory available on the target computing platform. In this section and the next, we focus on the adaptation of Myers-Miller algorithm.

It is easy to see that an optimal (global) alignment is comprised of an optimal alignment of $A[1 : m/2]$ and $B[1 : j]$ and an optimal alignment of $A[m : m/2 + 1]$ ($A[m : i]$ is the reverse of $A[i : m]$) and $B[n : j + 1]$ for some j, $1 \leq j \leq n$. The value of j for which this is true is called the *optimal crossover point*. Myers and Miller's linear space algorithm for alignment determines the optimal alignment by first determining the optimal crossover point (si, sj) where $si = m/2$, and then recursively aligning $A[1 : m/2]$ and $B[1 : sj]$ as well as $A[m : m/2 + 1]$ and $B[n : sj + 1]$. Equivalently, an optimal alignment of $A[1 : m]$ and $B[1 : n]$ is an optimal alignment of $A[1 : m/2]$ and $B[1 : sj]$ concatenated with the reverse of an optimal alignment of $A[m : m/2 + 1]$ and $B[n : sj + 1]$. Hence, an optimal alignment is comprised of a sequence of optimal crossover points. This is depicted visually in Figure 4. Figure 4(a) shows alignments using 3 possible crossover points at row $m/2$ of H. Figure 4(b) shows the partitioning of the alignment problem into 2 smaller alignment problems (shaded rectangles) using the optimal crossover point (meeting point of the 2 shaded rectangles) at row $m/2$. Figure 4(c) shows the partitioning of each of the 2 subproblems of Figure 4(b) using the optimal crossover points for these subproblems (note that these crossovers take place at rows $m/4$ and $3m/4$, respectively). Figure 4(d) shows the constructed optimal alignment, which is presently comprised of the 3 determined optimal crossover points.

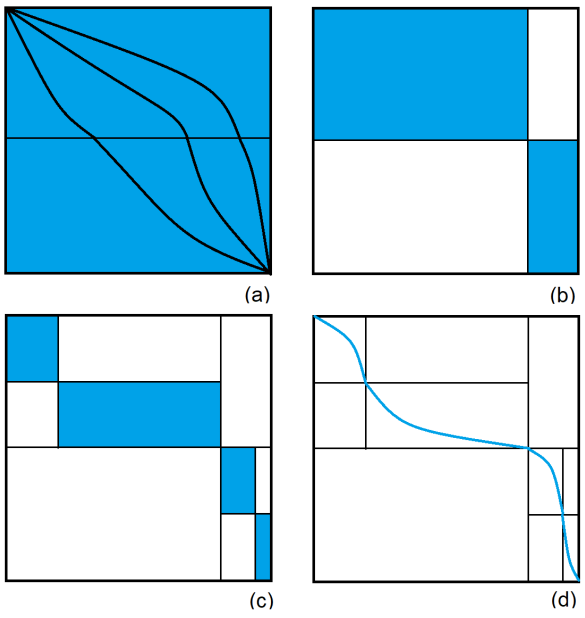


Figure 4: An alignment as crossover points. (a) Three alignments (b) Optimal crossover for m/2 (c)Optimal crossover points at m/4 and 3m/4 (d) Best alignment

Algorithm 4 Myers and Miller algorithm

```
1:  MM(A[1:m], B[1:n], path[1:m])
2:  if m <= 1 or n <= 1 then
3:      Linear search to find optimal crossover point [si, sj]
4:      path[m] ← [si, sj]
5:  else
6:      H_top ← MScore(A[1:m/2], B[1:n])
7:      H_bot ← MScore(A[m:m/2+1], B[n:1])
8:      Linear search to find optimal crossover point [si, sj]
9:      MM(A[1:m/2], B[1:sj], path[1:m/2])
10:     path[m/2] ← [si, sj]
11:     MM(A[m/2+1:m], B[sj+1:n], path[m/2+1:m])
12: end if
```

The Myers and Miller algorithm, MM (Algorithm 4), uses a modified version of the linear space scoring algorithm $Score$ (Algorithm 1) to obtain the scores for the best alignments of $A[1:i]$ and $B[1:j]$, $1 \le i \le m/2$, $1 \le j \le n$ as well as for the best alignments of $A[m:i]$ and $B[m:j]$, $m/2 < i \le m$, $1 \le j \le n$. This modified version $MScore$ differs from $Score$ only in that $MScore$ returns the entire array H rather than just $H[n]$. Using the returned H arrays for the forward and reverse alignments, the optimal crossover point for the best alignment is computed as in algorithm MM (Algorithm 4). Once the optimal crossover point is known, two recursive calls are made to optimally align the top and bottom halves of A with left and right parts of B. The approximately time complexity for iteration k is $O(2mn/2^k)$, hence the total time complexity is rough $2mn$.

In each level of recursion, the number of main memory accesses is dominated by those made in the calls to $MScore$. From the analysis for $Score$, it follows that when n is large, the number of accesses to main memory is $\approx 2mn/w(1+1/2+1/4+\cdots) \approx 4mn/w$.

3.2.2 Diagonal Myers and Miller Algorithm

Let $Mdiagonal$ be algorithm $Diagonal$ (Algorithm 2) modified to return the entire H array rather than just $H[n]$. Our diagonal Myers and Miller algorithm ($MMDiagonal$) replaces the two statements in algorithm MM (Algorithm 4) that invoke $MScore$ with a test that causes $Mdiagonal$ to be used in place of $MScore$ when both of m and n are sufficiently long.

From the analysis for $Diagonal$, it follows that when m and n are large, the number of accesses to main memory is $\approx 3mn/w(1+1/2+1/4+\cdots) \approx 6mn/w$.

3.2.3 Striped Myers and Miller Algorithm

Let $MStrip$ be algorithm $Strip$ (Algorithm 3) modified to return the entire H array rather than just $H[n]$. Our striped Myers and Miller algorithm ($MMStrip$) replaces the two statements in algorithm MM (Algorithm 4) that invoke $MScore$ with a test that causes $MStrip$ to be used in place of $MScore$ when both of m and n are sufficiently long.

From the analysis for $Strip$, it follows that when m and n are large, the number of accesses to main memory is $\approx 2mn/(wq)(1+1/2+1/4+\cdots) \approx 4mn/(wq)$.

3.3 Parallel Scoring Algorithms

3.3.1 Parallel Score Algorithm

As remarked earlier in $Score$ algorithm, the elements in a row of the score matrix need to be computed sequentially from left to right because of data dependencies. So, we are unable to parallelize the inner **for** loop of $Score$ (Algorithm 1). Instead, we adopt the unusual approach of parallelizing the outer **for** loop while computing the inner loop sequentially using a single processor. Initially, processor s is assigned to do the outer loop computation for $i = s$, $1 \le i \le p$, where p is the number of processors. Processor s begins after a suitable time lag relative to the start of processor $s-1$ so that the data it needs for its computation has already been computed by processor $s-1$. That is, processor 1 begins the inner loop computation for $i = 1$ at time 0, then, with a suitable time lag, processor 2 begins the outer loop computation for $i = 2$, then, with a further lag, processor 3 begins the $i = 3$ computation and so on. When a processor has finished with its iteration i computation, it starts on iteration $i+p$ of the outer loop. Synchronization primitives are used to ensure suitable time lags. The time complexity of the resulting p-core algorithm PP_Score is $O(mn/p)$.

3.3.2 Parallel Diagonal Algorithm

The inner **for** loop of $Diagonal$ (Algorithm 2) is easily parallelized as the elements on a diagonal are independent and may be computed simultaneously. So, in our parallel version, we divide the diagonal d into p blocks, where p is the number of processors. We assign a block to each processor from left to right as in Figure 5. The time complexity of the resulting p-core algorithm $PP_Diagonal$ is $O(mn/p)$.

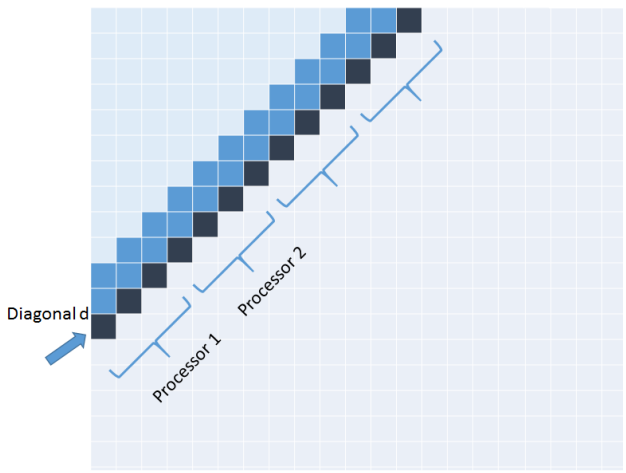


Figure 5: Parallel Diagonal algorithm.

3.3.3 Parallel Strip Algorithm

In the *Strip* scoring algorithm, we partition the score matrix H into n/q strips of size q (Figure 3) and compute the strips one at a time from left to right. Inside a strip, scores are computed row by row from top to bottom. We see that the computation of one strip can begin once the first row of the previous strip has been computed. In our parallel version of this algorithm, processor i is initially assigned to compute strip i, $1 \leq i \leq p$. When computing a value in its assigned strip, a processor needs to wait until the values (if any) needed from the strip to its left have been computed. When a processor completes the computation of strip j, it proceeds with the computation of strip $j+p$. Figure 6 shows a possible state in the described parallel strip computation strategy. We maintain an array $signal[]$ such that $signal[r] = s+1$ iff the row r computation for strips 1 through s has been completed. This array enables the processor working on the strip to its right to determine when it can begin the computation of its r'th row. The time complexity of the resulting parallel strip algorithm, PP_Strip, is $O(mn/p)$.

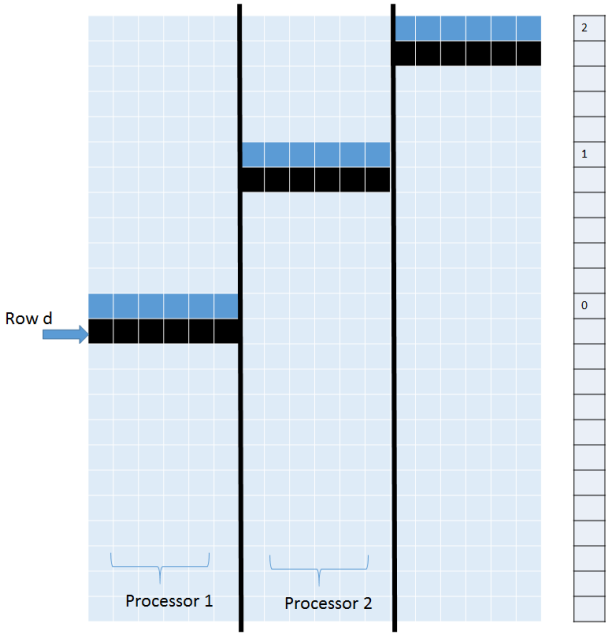


Figure 6: Parallel Strip algorithm.

3.4 Parallel Alignment Algorithms

In the single-core implementation, we divide the H matrix into two equal size parts and apply the scoring algorithm to each part. Then, we determine the optimal crossover point where the sum of the scores from both directions is maximum. This crossover point is used to divide the matrix into two smaller score matrices to which this decomposition strategy is recursively applied. The first application of this strategy yields two independent subproblems and following an application of the strategy to each of these subproblems, we have 4 even smaller subproblems. Following k rounds, we have 2^k independent subproblems.

For the parallel version of alignment algorithms, we employ the following strategies:

- When the number of independent matrices is small, each matrix is computed using the parallel version of score algorithms PP_Score, $PP_Diagonal$ and PP_Strip; where p processors are assigned to the parallel computation. In other words, the matrices are computed in sequence.

- When the number of independent matrices is large, each matrix is computed using the single-core algorithms $Score$, $Diagonal$ and $Strip$. Now, p matrices are concurrently computed.

Let PP_MM, $PP_MMDiagonal$ and $PP_MMStrip$, respectively, denote the parallel versions of MM, $MMDiagonal$ and $MMStrip$.

4 Results

4.1 Experimental Settings and Test Data

We implemented the single-core scoring and alignment algorithms in C and the multi-core scoring and alignment algorithms in C and OpenMP. The relative performance of these algorithms was measured on the following platforms:

1. Intel Xeon CPU E5-2603 v2 Quad-Core processor 1.8GHz with 10MB cache.

2. Intel I7-x980 Six-Core processor 3.33GHz with 12MB LLC cache.

3. Intel Xeon CPU E5-2695 v2 2xTwelve-Core processors 2.40GHz with 30MB cache.

For convenience, we will, at times, refer to these platforms as Xeon4, Xeon6, and Xeon24 (i.e., the number of cores is appended to the name Xeon).

All codes were compiled using the gcc compiler with the O2 option. On our Xeon4 platform, we used the "perf" [21] software to measure energy usage through the RAPL interface. So, for this platform, we report cache misses and energy consumption as well as running time. For the Xeon6 and Xeon24 platforms, we provide the running time only.

For test data, we used randomly generated protein sequences as well as real protein sequences obtained from the Globin Gene Server[22] and DNA/RNA/protein sequences from the National Center for Biotechnology Information (NCBI) database [23]. We used the BLOSUM62[1] scoring matrix for all our experiments. The results for our randomly generated protein sequences were comparable to those for similarly sized sequences used from the two databases [22] and [23]. So, we present only the results for the latter data sets here.

4.2 Xeon E5-2603 (Xeon4)

4.2.1 Score Algorithms

Figure 7 and Table 1 give the number of cache misses on our Xeon4 platform for different sequence sizes. The last two columns of Table 1 gives the percent reduction in the observed cache miss count of *Strip* relative to *Score* and *Diagonal*. *Strip* has the fewest cache misses followed by *Score* and *Diagonal* (in this order). *Strip* reduces cache misses by up to 86.2% relative to *Score* and by up to 92.3% relative to *Diagonal*.

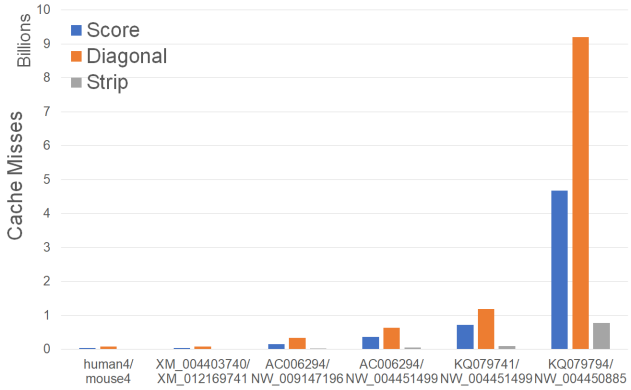


Figure 7: Cache misses of scoring algorithms, in billions, on Xeon4.

Figure 8 and Table 2 give the running times of our scoring algorithms on our Xeon4 platform. In the figure, the time is in seconds while in the table, the time is given using the format $hh : mm : ss$. The table also gives the percent reduction in running time achieved by *Strip* relative to *Score* and *Diagonal*.

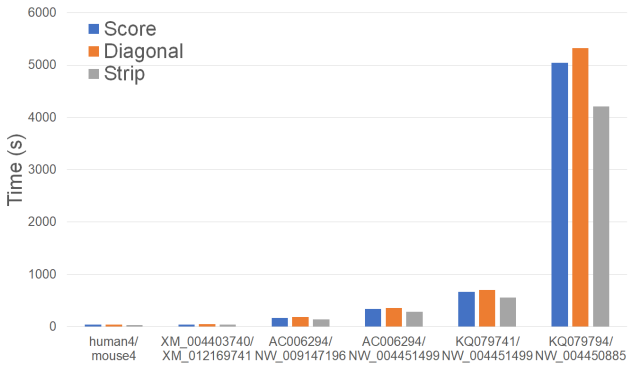


Figure 8: Run time of scoring algorithms, in seconds, on Xeon4.

As can be seen, on our Xeon4 platform, *Strip* is the fastest followed by *Score* and *Diagonal* (in this order). *Strip* reduces the running time by up to 17.5% relative to *Score* and by up to 22.8% relative to *Diagonal*. The reduction in running time, while significant, isn't as much as the reduction in cache misses possibly due to the effect of cache prefetching, which reduces cache induced computational delays.

Figures 9 and Tables 3 give the CPU and cache energy consumed, in joules, by our Xeon4 platform.

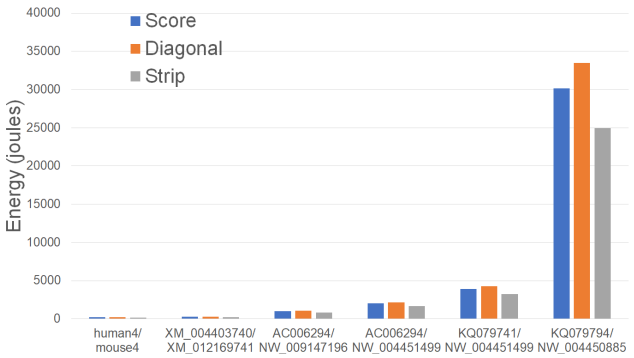


Figure 9: CPU and cache energy consumption of scoring algorithms, in joules, on Xeon4.

On our datasets, *Strip* required up to 18.5% less CPU and cache energy than *Score* and up to 25.5% less than *Diagonal*. It is interesting to note that the energy reduction is comparable to the reduction in running time suggesting a close relationship between running time and energy consumption for this application.

4.2.2 Parallel Scoring Algorithms

Figure 10 and Table 4 give the number of cache misses on our Xeon4 platform for our parallel scoring algorithms. *PP_Strip* has the fewest cache misses followed by *PP_Score* and *PP_Diagonal* (in this order). *PP_Strip* reduces cache misses by up to 98.1% relative to *PP_Score* and by up to 99.1% relative to *PP_Diagonal*. We observe also that the total cache misses for *PP_Score* is slightly higher than for *Score* for smaller instances and lower for larger instances. *PP_Diagonal*, on the other hand, consistently has more cache misses than *Diagonal*. *PP_Strip* exhibits a significant reduction in cache misses. This is because we chose the strip width to be such that p strip rows fit in this cache. Most of the cache misses in the *Strip* are from the vector that transfers boundary results from one strip to the next. When p strips are being worked on simultaneously, the inter-strip data that is to be transferred is often in the cache and so many of the cache misses incurred by the single-core algorithm are saved. The remaining two algorithms do not allow this flexibility in choosing the segment size a processor works on; this size is fixed at $O(n/p)$.

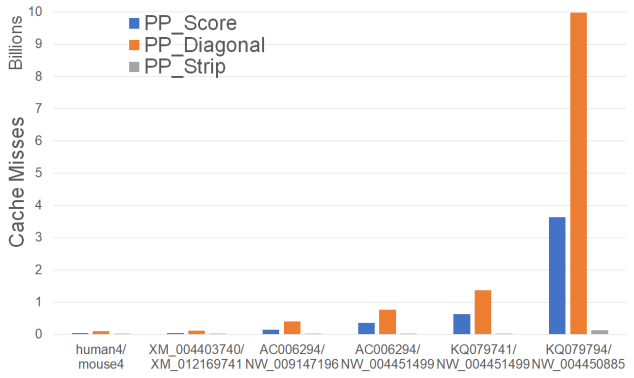


Figure 10: Cache misses of parallel scoring algorithms, in billions, on Xeon4.

Table 1: Cache misses of scoring algorithms, in millions, on Xeon4.

A	|A|	B	|B|	Score	Diagonal	Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	33	74	6	82.1%	92.1%
XM_004403740	104,267	XM_012169741	103,004	38	86	7	81.8%	91.9%
AC006294	200,000	NW_009147196	200,000	144	342	26	81.9%	92.3%
AC006294	200,000	NW_004451499	398,273	363	630	52	85.6%	91.7%
KQ079741	392,981	NW_004451499	398,273	712	1,190	98	86.2%	91.7%
KQ079794	1,083,068	NW_004450885	1,098,196	4,676	9,200	772	83.5%	91.6%

Table 2: Run time of scoring algorithms, in hh:mm:ss, on Xeon4.

A	|A|	B	|B|	Score	Diagonal	Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:39	0:00:42	0:00:33	16.2%	21.6%
XM_004403740	104,267	XM_012169741	103,004	0:00:45	0:00:49	0:00:38	16.2%	22.8%
AC006294	200,000	NW_009147196	200,000	0:02:49	0:03:02	0:02:22	16.2%	22.1%
AC006294	200,000	NW_004451499	398,273	0:05:42	0:05:58	0:04:42	17.5%	21.3%
KQ079741	392,981	NW_004451499	398,273	0:11:03	0:11:44	0:09:14	16.4%	21.4%
KQ079794	1,083,068	NW_004450885	1,098,196	1:24:06	1:28:45	1:10:10	16.6%	20.9%

Figure 11 and Table 5 give the running times for our parallel scoring algorithms on our Xeon4 platform. In the figure, the time is in seconds while in the table, the time is given using the format $hh:mm:ss$. As in the table, *PP_Strip* is the fastest algorithm in practice, which is up to 40.0% faster than *PP_Score* and up to 38.4% faster than *PP_Diagonal*.

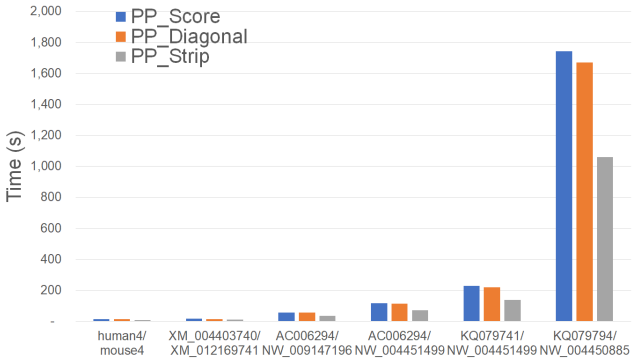


Figure 11: Run time of parallel scoring algorithms, in seconds, on Xeon4.

Table 6 gives the speedup of each of our parallel scoring algorithms relative to their sequential counterparts. As can be seen, the speedup of *PP_Strip* (i.e., *Strip/PP_Strip*) is between 3.92 and 3.98, which is quite close to the number of cores (4) on our Xeon4 platform. *PP_Score* achieves a speedup in the range 2.82 to 2.94 and the speedup for *PP_Diagonal* is in the range 3.12 to 3.21.

The excellent speedup exhibited by *PP_Strip* is due largely to our ability to greatly reduce cache misses for this algorithm.

Figures 12 and Tables 7 give the CPU and cache energy consumed, in joules, by our Xeon4 platform. On our datasets, *PP_Strip* required up to 41.2% less CPU and cache energy than *PP_Score* and up to 45.5% less than *PP_Diagonal*.

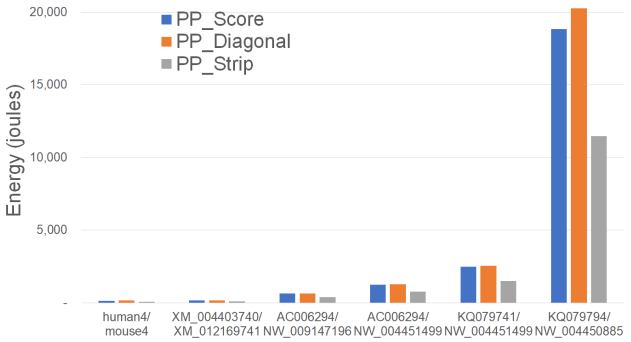


Figure 12: CPU and cache energy consumption of parallel scoring algorithms, in joules, on Xeon4.

Compared to the sequential scoring algorithms, the multi-core algorithms use higher CPU power but less running time. Since the power increase is less than the decrease in running time, energy consumption is reduced.

4.2.3 Alignment Algorithms

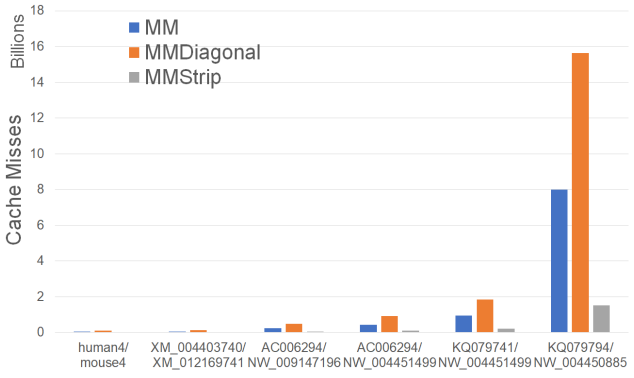


Figure 13: Cache misses for alignment algorithms, in billions, on Xeon4.

Figure 13 and Table 8 give the number of cache misses of our single-core alignment algorithms on our Xeon4 plat-

Table 3: CPU and cache energy consumption of scoring algorithms, in joules, on Xeon4.

A	|A|	B	|B|	Score	Diagonal	Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	230.67	252.29	190.6	17.4%	24.5%
XM_004403740	104,267	XM_012169741	103,004	268.29	297.28	221.49	17.4%	25.5%
AC006294	200,000	NW_009147196	200,000	997.88	1100.6	829.36	16.9%	24.6%
AC006294	200,000	NW_004451499	398,273	2026.26	2178.97	1651.46	18.5%	24.2%
KQ079741	392,981	NW_004451499	398,273	3944.3	4300.59	3253.46	17.5%	24.3%
KQ079794	1,083,068	NW_004450885	1,098,196	30125.93	33472.81	24980.35	17.1%	25.4%

Table 4: Cache misses of parallel scoring algorithms, in millions, on Xeon4.

A	|A|	B	|B|	PP_Score	PP_Diagonal	PP_Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	34	102	1	96.7%	98.9%
XM_004403740	104,267	XM_012169741	103,004	39	115	1	96.8%	98.9%
AC006294	200,000	NW_009147196	200,000	146	398	4	96.9%	98.9%
AC006294	200,000	NW_004451499	398,273	362	768	7	98.1%	99.1%
KQ079741	392,981	NW_004451499	398,273	628	1,373	14	97.7%	99.0%
KQ079794	1,083,068	NW_004450885	1,098,196	3,642	9,976	121	96.7%	98.8%

form. *MMStrip* has the fewest number of cache misses followed by *MM* and *MMDiagonal* (in this order). *MMStrip* reduces cache misses by up to 81.0% relative to *MM* and by up to 90.3% relative to *MMDiagonal*.

Figure 14 and Table 9 give the running times of our single-core alignment algorithms on our Xeon4 platform. As can be seen, *MMStrip* is the fastest followed by *MM* and *MMDiagonal* (in this order). *MMStrip* reduces running time by up to 15.0% relative to *MM*, by up to 13.4% relative to *MMDiagonal*. As was the case with our scoring algorithms, the reduction in running time, while significant, isn't as much as the reduction in cache misses.

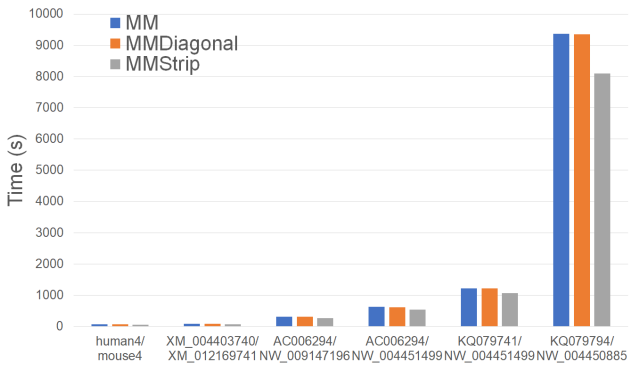


Figure 14: Run time of alignment algorithms, in Seconds, on Xeon4.

Figures 15 and Tables 10 give the CPU and cache energy consumption, in joules, by our single-core alignment algorithms. On our datasets, *MMStrip* reduced up to 17.5% less CPU and cache energy than *MM* and up to 18.7% less than *MMDiagonal*. Once again, the energy reduction is comparable to the reduction in running time suggesting a close relationship between running time and energy consumption for this application.

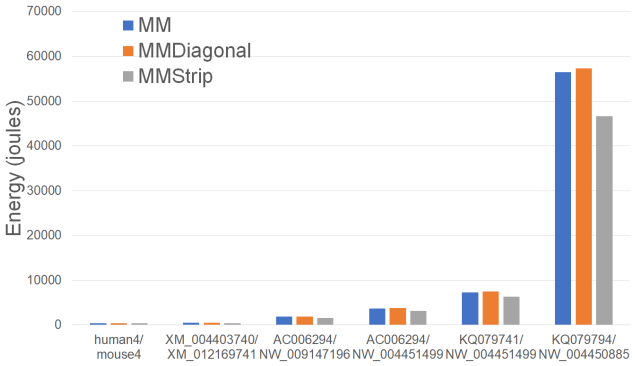


Figure 15: CPU and cache energy consumption of alignment algorithms, in joules, on Xeon4.

4.2.4 Parallel Alignment Algorithms

Figure 16 and Table 11 give the number of cache misses of our multi-core alignment algorithms on our Xeon4 platform.

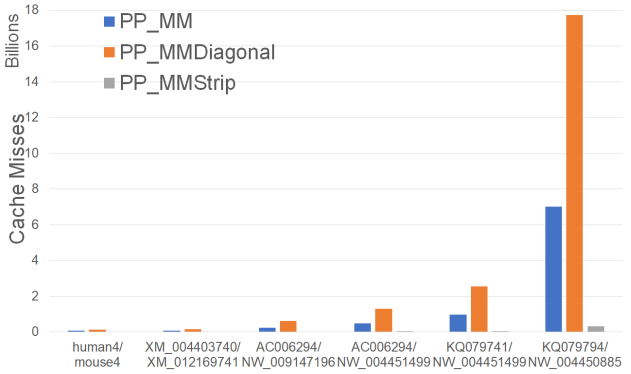


Figure 16: Cache misses for parallel alignment algorithms, in billions, on Xeon4.

PP_MMStrip has the fewest number of cache misses followed by *PP_MM* and *PP_MMDiagonal* (in this order). *PP_MMStrip* reduces cache misses by

Table 5: Run time of parallel scoring algorithms on Xeon4.

A	|A|	B	|B|	PP_Score	PP_Diagonal	PP_Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:14	0:00:13	0:00:08	40.0%	37.7%
XM_004403740	104,267	XM_012169741	103,004	0:00:16	0:00:15	0:00:10	39.8%	37.1%
AC006294	200,000	NW_009147196	200,000	0:00:58	0:00:58	0:00:36	39.0%	38.4%
AC006294	200,000	NW_004451499	398,273	0:01:56	0:01:53	0:01:11	39.0%	37.0%
KQ079741	392,981	NW_004451499	398,273	0:03:49	0:03:39	0:02:19	39.3%	36.6%
KQ079794	1,083,068	NW_004450885	1,098,196	0:29:02	0:27:52	0:17:39	39.2%	36.7%

Table 6: Speedup of parallel scoring algorithms on Xeon4.

A	|A|	B	|B|	Score/PP	Diagonal/PP	Strip/PP
human4	97,634	mouse4	94,647	2.82	3.12	3.93
XM_004403740	104,267	XM_012169741	103,004	2.82	3.19	3.92
AC006294	200,000	NW_009147196	200,000	2.89	3.14	3.97
AC006294	200,000	NW_004451499	398,273	2.94	3.18	3.97
KQ079741	392,981	NW_004451499	398,273	2.89	3.21	3.98
KQ079794	1,083,068	NW_004450885	1,098,196	2.90	3.18	3.98

up to 95.5% relative to *PP_MM* and by up to 98.2% relative to *PP_MMDiagonal*.

Figure 17 and Table 12 give the running times for our parallel alignment algorithms on the Xeon4 platform. *PP_MMStrip* is faster than *PP_MM* by up to 37.4% and faster than *PP_MMDiagonal* by up to 40.3%.

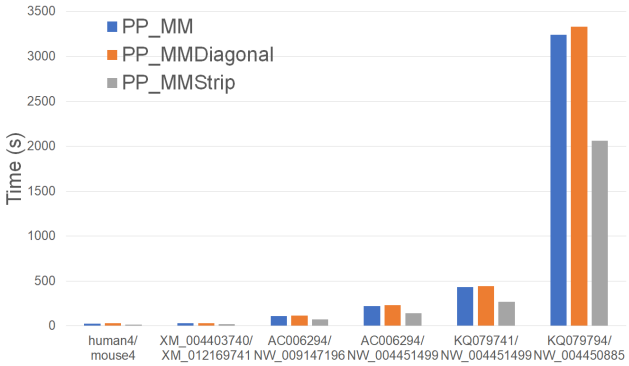


Figure 17: Run time of parallel alignment algorithms, in seconds, on Xeon4.

Table 13 gives the speedup of each parallel alignment algorithm relative to its single-core counterpart. The speedup achieved by *PP_MMStrip* (relative to *MMStrip*) ranges from 3.56 to 3.94 while that for *PP_MM* is in the range 2.77 to 2.88 and that for *PP_MMDiagonal* is in the range 2.53 to 2.81.

Figures 18 and Tables 14 give the CPU and cache energy consumption, in joules, by our multi-core alignment algorithms. On our datasets, *PP_MMStrip* required up to 29.9% less CPU and cache energy than *PP_MM* and up to 42.1% less than *PP_MMDiagonal*. Once again, the energy reduction is comparable to the reduction in running time suggesting a close relationship between running time and energy consumption for this application.

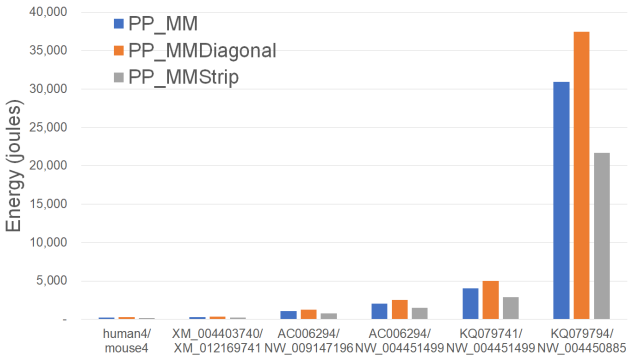


Figure 18: CPU and cache energy consumption of parallel alignment algorithms, in joules, on Xeon4.

4.3 I7-x980 (Xeon6)

4.3.1 Scoring Algorithms

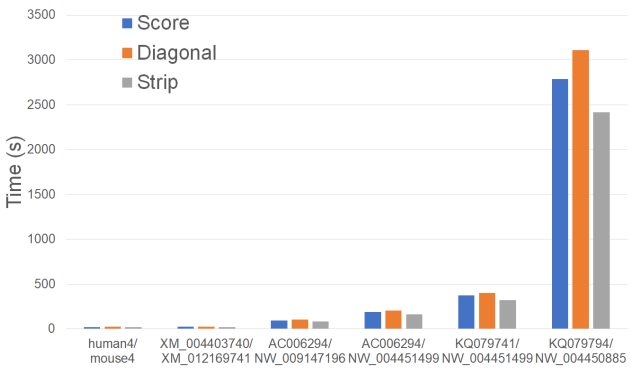


Figure 19: Run time of scoring algorithms, in seconds, on Xeon6

Figure 19 and Table 15 give the running times of our single-core scoring algorithms on our Xeon6 platform. As can be seen, *Strip* is the fastest followed by *Score* and *Diagonal* (in this order). *Strip* reduces running time by up to 14.3% relative to *Score* and by up to 22.4% relative to *Diagonal*.

Table 7: CPU and cache energy consumption of parallel scoring algorithms on Xeon4.

A	|A|	B	|B|	PP_Score	PP_Diagonal	PP_Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	150.12	161.77	88.23	41.2%	45.5%
XM_004403740	104,267	XM_012169741	103,004	174.30	175.68	102.76	41.0%	41.5%
AC006294	200,000	NW_009147196	200,000	631.46	661.29	381.02	39.7%	42.4%
AC006294	200,000	NW_004451499	398,273	1,256.11	1,294.55	760.57	39.5%	41.2%
KQ079741	392,981	NW_004451499	398,273	2,471.51	2,535.50	1,493.38	39.6%	41.1%
KQ079794	1,083,068	NW_004450885	1,098,196	18,830.03	20,253.39	11,477.30	39.0%	43.3%

Table 8: Cache misses for alignment algorithms, in millions, on Xeon4.

A	|A|	B	|B|	MM	MMDiagonal	MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	42	96	12	71.6%	87.6%
XM_004403740	104,267	XM_012169741	103,004	54	119	14	74.5%	88.5%
AC006294	200,000	NW_009147196	200,000	227	488	48	78.6%	90.1%
AC006294	200,000	NW_004451499	398,273	441	926	95	78.4%	89.7%
KQ079741	392,981	NW_004451499	398,273	959	1,851	200	79.1%	89.2%
KQ079794	1,083,068	NW_004450885	1,098,196	7,994	15,642	1,521	81.0%	90.3%

4.3.2 Parallel Scoring Algorithms

Figure 20 and Table 16 give the running times for our parallel scoring algorithms on our Xeon6 platform. As with *Xeon*4, *PP_Strip* is faster than *PP_Score* and *PP_Diagonal* and reduces the running time by up to 42.5% and 55.6%, respectively.

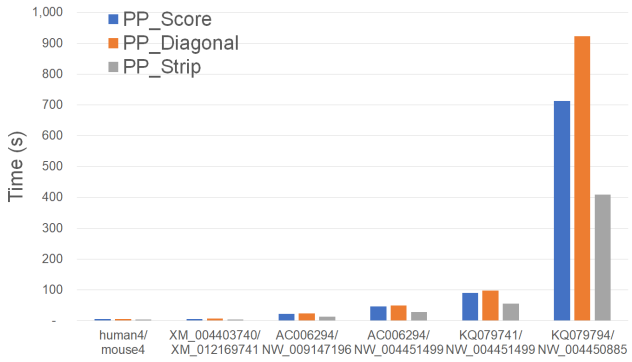


Figure 20: Run time of parallel scoring algorithms, in seconds, on Xeon6.

Table 18 gives the speedup of each of our parallel algorithms relative to their single-core counterparts. *PP_Strip* achieves a speedup of up to 5.89, which is very close to the number of cores. The maximum speedup achieved by *PP_Score* and *PP_Diagonal* was 4.09 and 4.25, respectively.

4.3.3 Alignment Algorithms

Figure 21 and Table 17 give the running times of our parallel scoring algorithms on the Xeon6 platform. As can be seen, *MMStrip* is the fastest followed by *MM* and *MMDiagonal* (in this order). *MMStrip* reduces running time by up to 12.6% relative to *MM* and by up to 14.2% relative to *MMDiagonal*.

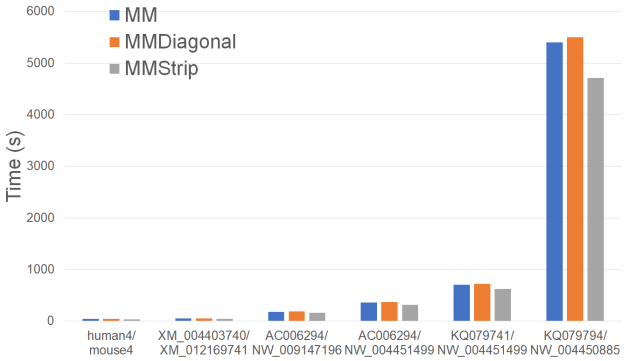


Figure 21: Run time of alignment algorithms, in seconds, on Xeon6.

4.3.4 Parallel Alignment Algorithms

Figure 22 and Table 20 give the running times of our parallel alignment algorithms on the Xeon6. *PP_MMStrip* is faster than *PP_MM* and *PP_MMDiagonal* and reduces the running time by up to 39.9% and 44.8%, respectively.

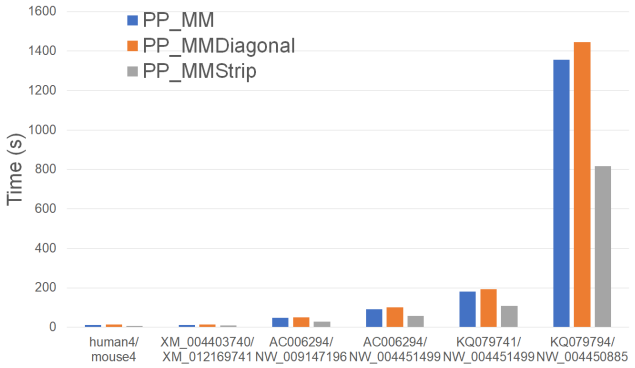


Figure 22: Run time of parallel alignment algorithms, in seconds, on Xeon6.

Table 21 gives the speedup of each of our parallel algorithms relative to their single-core counter-

Table 9: Run time of alignment algorithms on Xeon4.

A	|A|	B	|B|	MM	MMDiagonal	MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:01:12	0:01:13	0:01:03	12.6%	12.7%
XM_004403740	104,267	XM_012169741	103,004	0:01:27	0:01:24	0:01:14	15.0%	12.8%
AC006294	200,000	NW_009147196	200,000	0:05:16	0:05:14	0:04:33	13.6%	13.0%
AC006294	200,000	NW_004451499	398,273	0:10:25	0:10:18	0:09:04	13.0%	12.0%
KQ079741	392,981	NW_004451499	398,273	0:20:27	0:20:28	0:17:47	13.0%	13.1%
KQ079794	1,083,068	NW_004450885	1,098,196	2:36:12	2:35:53	2:14:59	13.6%	13.4%

Table 10: CPU and cache energy consumption of alignment algorithms on Xeon4.

A	|A|	B	|B|	MM	MMDiagonal	MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	426.38	438.07	369.94	13.2%	15.6%
XM_004403740	104,267	XM_012169741	103,004	509.66	511.32	431.59	15.3%	15.6%
AC006294	200,000	NW_009147196	200,000	1873.99	1911.35	1600.77	14.6%	16.2%
AC006294	200,000	NW_004451499	398,273	3707.55	3772.38	3189.38	14.0%	15.5%
KQ079741	392,981	NW_004451499	398,273	7313.49	7512.26	6278.63	14.2%	16.4%
KQ079794	1,083,068	NW_004450885	1,098,196	56478.59	57339.29	46589.55	17.5%	18.7%

parts. *PP_MMStrip* achieves a speedup of up to 5.78, which is very close to the number of cores. The maximum speedup achieved by *PP_MM* and *PP_MMDiagonal* was 3.98 and 3.80, respectively.

4.4 Xeon E5-2695 (Xeon24)

4.4.1 Scoring Algorithms

Figure 23 and Table 19 give the running times of our single-core scoring algorithms on our Xeon24 platform. As was the case on our other test platforms, *Strip* is the fastest followed by *Score* and *Diagonal* (in this order). *Strip* reduces running time by up to 19.7% relative to *Score* and by up to 35.1% relative to *Diagonal*.

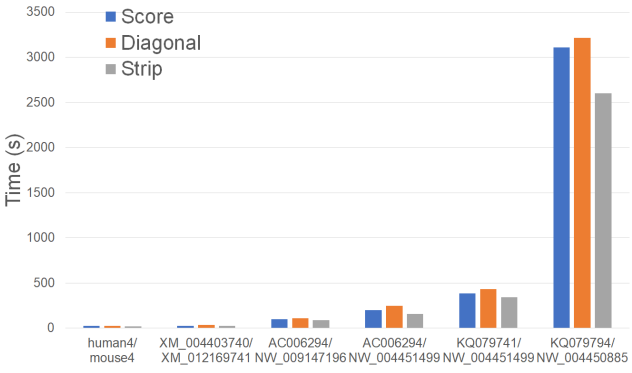


Figure 23: Run time of scoring algorithms, in seconds, on Xeon24.

4.4.2 Parallel Scoring Algorithms

Figure 24 and Table 22 give the running times for our parallel scoring algorithms on our Xeon24 platform. *PP_Strip* is faster than *PP_Score* and *PP_Diagonal* and reduces the running time by up to 61.4% and 76.2%, respectively.

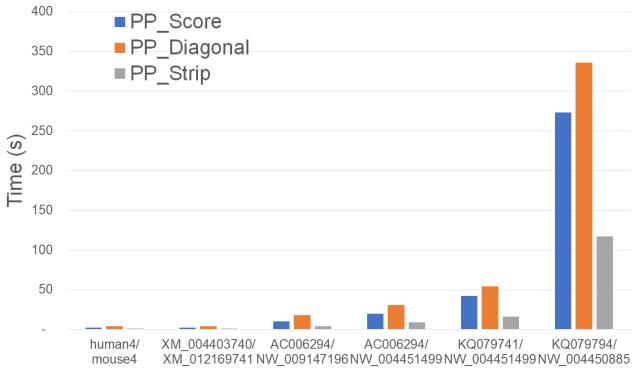


Figure 24: Run time of parallel scoring algorithms, in seconds, on Xeon24.

Table 23 gives the achieved speedup. *PP_Strip* scales quite well and results in a speedup of up to 22.22. The maximum speedups provided by *PP_Score* and *PP_Diagonal* are 11.36 and 9.56, respectively.

4.4.3 Alignment Algorithms

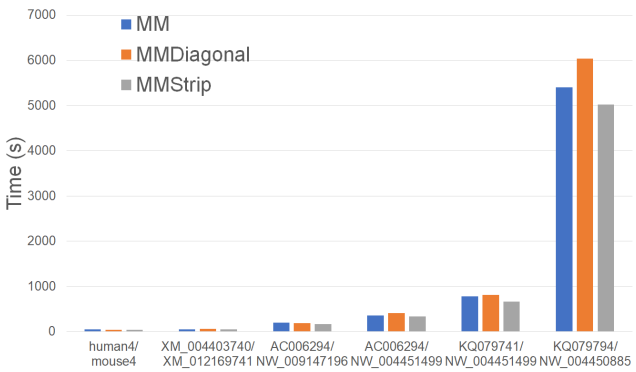


Figure 25: Run time of alignment algorithms, in seconds, on Xeon24.

Figure 25 and Table 24 give the running times of our single-core alignment algorithms on our Xeon24 plat-

Table 11: Cache misses for parallel alignment algorithms, in millions, on Xeon4.

A	|A|	B	|B|	PP_MM	PP_MMDiagonal	PP_MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	64	137	9	85.3%	93.1%
XM_004403740	104,267	XM_012169741	103,004	71	161	17	75.8%	89.4%
AC006294	200,000	NW_009147196	200,000	251	625	23	91.0%	96.4%
AC006294	200,000	NW_004451499	398,273	495	1,296	39	92.2%	97.0%
KQ079741	392,981	NW_004451499	398,273	975	2,558	52	94.6%	98.0%
KQ079794	1,083,068	NW_004450885	1,098,196	7,001	17,721	314	95.5%	98.2%

Table 12: Run time of parallel alignment algorithms on Xeon4.

A	|A|	B	|B|	PP_MM	PP_MMDiagonal	PP_MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:26	0:00:29	0:00:17	34.5%	40.3%
XM_004403740	104,267	XM_012169741	103,004	0:00:30	0:00:33	0:00:21	31.6%	37.3%
AC006294	200,000	NW_009147196	200,000	0:01:52	0:01:56	0:01:12	36.1%	38.5%
AC006294	200,000	NW_004451499	398,273	0:03:43	0:03:50	0:02:20	37.4%	39.1%
KQ079741	392,981	NW_004451499	398,273	0:07:12	0:07:23	0:04:31	37.2%	38.8%
KQ079794	1,083,068	NW_004450885	1,098,196	0:54:03	0:55:33	0:34:20	36.5%	38.2%

form. *MMStrip* is the fastest followed by *MM* and *MMDiagonal* (in this order). *MMStrip* reduces running time by up to 17.1% relative to *MM* and by up to 16.8% relative to *MMDiagonal*.

4.4.4 Parallel Alignment Algorithms

Figure 26 and Table 25 give the running times of our parallel alignment algorithms on Xeon24. As can be seen, *PP_MMStrip* is faster than *PP_MM* and *PP_MMDiagonal*. It reduces the running time by up to 47.3% and 84.6%, respectively.

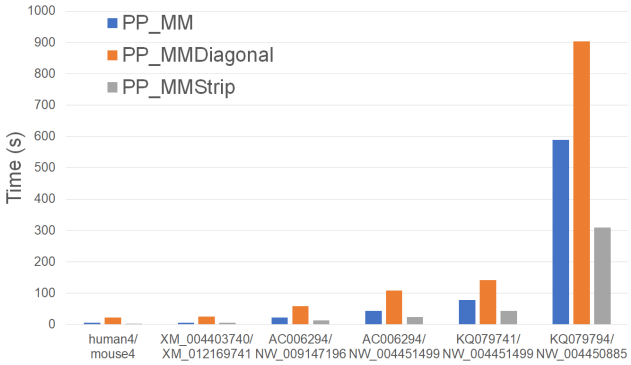


Figure 26: Run time of parallel alignment algorithms, in seconds, on Xeon24.

Table 26 gives the speedup of our parallel algorithms. *PP_MMStrip* achieves a speedup of up to 16.2 while *PP_MM* and *PP_MMDiagonal* have maximum speedups of 9.79 and 6.58.

5 Discussion

By accounting for the presence of caches in modern computers, we are able to arrive at sequence alignment algorithms that are considerably faster than those that do not take advantage of computer caches. Our benchmarking demonstrates the value of optimizing cache usage. Our cache-efficient algorithms *Strip* and *MMStrip* were the best-performing single-core algorithms and their parallel counterparts were the best-performing parallel algorithms. *Strip* reduced running time by as much as 19.7% relative to the classical scoring algorithm *Score* due to Smith and Waterman and *MM_Strip* reduced running time by as much as 17.1% relative to the alignment algorithm of Myers and Miller. Neither the algorithm of Smith and Waterman nor that of Myers and Miller optimize cache utilization. The parallel versions of *Strip* and *MM_Strip* were up to 61.4% and 47.3% faster than the parallel versions of the Smith and Waterman and the Myers and Miller algorithms, respectively.

6 Limitations

Our cache miss analyses assume a simple cache model in which there is a single LRU cache. In practice, computers have multiple levels of cache and employ sophisticated and proprietary cache replacement strategies. Despite the use of a simplified cache model for analysis, the developed cache-efficient algorithms perform very well in practice.

7 Conclusion

The main contributions of this papers are

1. cache efficient single-core and multi-core algorithms to determine the score of the best alignment;
2. cache efficient single-core and multi-core algorithms to determine the best alignment.

The effectiveness of our cache-efficient algorithms has been demonstrated experimentally using three computational platforms. Future work includes developing the cache-efficient algorithms for other problems in computational biology.

Conflict of Interest The authors declare no conflict of interest.

Table 13: Speedup of parallel alignment algorithms on Xeon4.

A	|A|	B	|B|	MM/PP	MMDiagonal/PP	MMStrip/PP
human4	97,634	mouse4	94,647	2.77	2.53	3.70
XM_004403740	104,267	XM_012169741	103,004	2.78	2.56	3.56
AC006294	200,000	NW_009147196	200,000	2.80	2.70	3.81
AC006294	200,000	NW_004451499	398,273	2.84	2.69	3.89
KQ079741	392,981	NW_004451499	398,273	2.84	2.78	3.94
KQ079794	1,083,068	NW_004450885	1,098,196	2.88	2.81	3.93

Table 14: CPU and cache energy consumption of parallel alignment algorithms on Xeon4.

A	|A|	B	|B|	PP_MM	PP_MMDiagonal	PP_MMStrip	Imp1	Imp2
human4	97634	mouse4	94647	236.98	305.14	181.6	23.4%	40.5%
XM_004403740	104267	XM_012169741	103004	272.26	352.56	216.83	20.4%	38.5%
AC006294	200000	NW_009147196	200000	1044.86	1279.77	747.86	28.4%	41.6%
AC006294	200000	NW_004451499	398273	2039.55	2540.52	1483.8	27.2%	41.6%
KQ079741	392981	NW_004451499	398273	4020.48	4979.17	2905.32	27.7%	41.7%
KQ079794	1083068	NW_004450885	1098196	30964.02	37461.35	21703.65	29.9%	42.1%

Acknowledgment This work was supported, in part, by the National Science Foundation under award NSF 1447711.

References

[1] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc Natl Acad Sci U S A*, vol. 89, pp. 10 915–10 919, 1992.

[2] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443–453, 1970.

[3] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[4] O. Gotoh, "An improved algorithm for matching biological sequences," *Journal of Molecular Biology*, vol. 162, pp. 705–708, 1982.

[5] D. S. Hirschberg, "A linear space algorithm for computing longest common subsequences," *Communications of the ACM*, vol. 18, pp. 341–343, 1975.

[6] E. Myers and W. Miller, "Optimal alignments in linear space," *Computer Applications in the Biosciences(CABIOS)*, vol. 4, pp. 11–17, 1988.

[7] X. Huang, R. Hardison, and W. Miller, "A space-efficient algorithm for local similarities," *Comput Appl Biosci*, vol. 6, p. 373âĂŞ381, 1990.

[8] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[9] W. Pearson and D. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences USA*, vol. 85, pp. 2444–2448, 1988.

[10] K. Chao, J. Zhang, J. Ostell, and W. Miller, "A local alignment tool for very long dna sequences," *Comput Appl Biosci*, vol. 11, pp. 147–153, 1995.

[11] A. Driga, P. Lu, J. Schaeffer, D. Szafron, K. Charter, and I. Parsons, "Fastlsa: a fast, linear-space, parallel and sequential algorithm for sequence alignment," *Algorithmica*, vol. 45, p. 337âĂŞ375, 2006.

[12] J. Li, S. Ranka, and S. Sahni, "Pairwise sequence alignment for very long sequences on gpus," *IEEE 2nd International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 2012.

[13] E. O. Sandes and A. C. M. A. Melo, "Smith-waterman alignment of huge sequences with gpu in linear space," *IEEE International Symposium on Parallel and Distributed Processing (IPDPS)*, pp. 1199–1211, 2011.

[14] S. Aluru and N. Jammula, "A review of hardware acceleration for computational genomics," *IEEE Design and Test*, vol. 31, pp. 19–30, 2014.

[15] S. Rajko and S. Aluru, "Space and time optimal parallel sequence alignments," *IEEETPDS:IEEE Transactions on Parallel and Distributed Systems*, vol. 15, 2004.

[16] A. Khajeh-Saeed, S. Poole, and J. B. Perot, "Acceleration of the smithâĂŞwaterman algorithm using single and multiple graphics processors," *Journal of Computational Physics*, p. 4247âĂŞ4258, 2010.

[17] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," *Genome Biol*, vol. 5, 2004.

[18] T. Almeida and N. Roma, "A parallel programming framework for multi-core dna sequence alignment," *Complex, Intelligent and Software Intensive Systems (CISIS), 2010 International Conference on*, pp. 907 – 912, 2010.

[19] K. Hamidouche, F. M. Mendonca, J. Falcou, A. C. M. A. Melo, and D. Etiemble, "Parallel smith-waterman comparison on multicore and manycore computing platforms with bsp++," *International Journal of Parallel Programming*, vol. 41, pp. 1110–136, 2013.

[20] C. Zhao and S. Sahni, "Cache and energy efficient alignment of very long sequences," *2015 IEEE 5th international conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, 2015.

[21] "Perf tool," `https://perf.wiki.kernel.org/index.php/Main_Page`.

[22] "Globin gene server," `http://globin.cse.psu.edu/globin/html/pip/examples.html`.

[23] "Ncbi database," `http://www.ncbi.nlm.nih.gov/gquery`.

Table 15: Run time of scoring algorithms, in hh:mm:ss, on Xeon6.

A	|A|	B	|B|	Score	Diagonal	Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:21	0:00:24	0:00:19	13.2%	21.1%
XM_004403740	104,267	XM_012169741	103,004	0:00:25	0:00:27	0:00:22	13.2%	20.2%
AC006294	200,000	NW_009147196	200,000	0:01:33	0:01:43	0:01:21	13.2%	21.7%
AC006294	200,000	NW_004451499	398,273	0:03:08	0:03:25	0:02:41	14.3%	21.4%
KQ079741	392,981	NW_004451499	398,273	0:06:13	0:06:43	0:05:19	14.3%	20.8%
KQ079794	1,083,068	NW_004450885	1,098,196	0:46:27	0:51:51	0:40:16	13.3%	22.4%

Table 16: Run time of parallel scoring algorithms, in hh:mm:ss, on Xeon6.

A	|A|	B	|B|	PP_Score	PP_Diagonal	PP_Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:06	0:00:06	0:00:04	27.9%	30.4%
XM_004403740	104,267	XM_012169741	103,004	0:00:06	0:00:07	0:00:04	37.4%	40.6%
AC006294	200,000	NW_009147196	200,000	0:00:23	0:00:24	0:00:14	38.8%	41.2%
AC006294	200,000	NW_004451499	398,273	0:00:47	0:00:49	0:00:28	40.6%	43.0%
KQ079741	392,981	NW_004451499	398,273	0:01:31	0:01:38	0:00:55	39.0%	43.4%
KQ079794	1,083,068	NW_004450885	1,098,196	0:11:53	0:15:23	0:06:50	42.5%	55.6%

Table 17: Run time of alignment algorithms, in hh:mm:ss, on Xeon6.

A	|A|	B	|B|	MM	MMDiagonal	MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:41	0:00:42	0:00:37	11.2%	12.7%
XM_004403740	104,267	XM_012169741	103,004	0:00:48	0:00:49	0:00:43	11.3%	12.8%
AC006294	200,000	NW_009147196	200,000	0:03:00	0:03:05	0:02:39	11.7%	14.0%
AC006294	200,000	NW_004451499	398,273	0:05:59	0:06:06	0:05:16	11.8%	13.7%
KQ079741	392,981	NW_004451499	398,273	0:11:45	0:12:01	0:10:22	11.8%	13.7%
KQ079794	1,083,068	NW_004450885	1,098,196	1:29:59	1:31:40	1:18:37	12.6%	14.2%

Table 18: Speedup of parallel scoring algorithms on Xeon6.

A	|A|	B	|B|	Score/PP	Diagonal/PP	Strip/PP
human4	97,634	mouse4	94,647	3.91	4.15	4.70
XM_004403740	104,267	XM_012169741	103,004	3.99	4.12	5.54
AC006294	200,000	NW_009147196	200,000	3.99	4.25	5.66
AC006294	200,000	NW_004451499	398,273	3.98	4.17	5.74
KQ079741	392,981	NW_004451499	398,273	4.09	4.11	5.76
KQ079794	1,083,068	NW_004450885	1,098,196	3.91	3.37	5.89

Table 19: Run time of scoring algorithms, in hh:mm:ss, on Xeon24.

A	|A|	B	|B|	Score	Diagonal	Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:23	0:00:26	0:00:20	12.1%	20.9%
XM_004403740	104,267	XM_012169741	103,004	0:00:26	0:00:35	0:00:24	8.9%	32.9%
AC006294	200,000	NW_009147196	200,000	0:01:41	0:01:51	0:01:28	13.5%	21.2%
AC006294	200,000	NW_004451499	398,273	0:03:18	0:04:05	0:02:39	19.7%	35.1%
KQ079741	392,981	NW_004451499	398,273	0:06:23	0:07:14	0:05:44	10.2%	20.7%
KQ079794	1,083,068	NW_004450885	1,098,196	0:51:48	0:53:34	0:43:19	16.4%	19.1%

Table 20: Run time of parallel alignment algorithms on Xeon6.

A	|A|	B	|B|	PP_MM	PP_MMDiagonal	PP_MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:11	0:00:13	0:00:07	33.6%	44.8%
XM_004403740	104,267	XM_012169741	103,004	0:00:12	0:00:15	0:00:08	33.4%	44.2%
AC006294	200,000	NW_009147196	200,000	0:00:48	0:00:51	0:00:29	39.2%	42.9%
AC006294	200,000	NW_004451499	398,273	0:01:32	0:01:42	0:00:58	37.5%	43.7%
KQ079741	392,981	NW_004451499	398,273	0:03:01	0:03:14	0:01:49	39.7%	43.8%
KQ079794	1,083,068	NW_004450885	1,098,196	0:22:37	0:24:06	0:13:36	39.9%	43.6%

Table 21: Speedup of parallel alignment algorithms on Xeon6.

A	|A|	B	|B|	MM/PP	MMDiagonal/PP	MMStrip/PP
human4	97,634	mouse4	94,647	3.81	3.22	5.10
XM_004403740	104,267	XM_012169741	103,004	3.88	3.31	5.17
AC006294	200,000	NW_009147196	200,000	3.72	3.59	5.41
AC006294	200,000	NW_004451499	398,273	3.88	3.58	5.48
KQ079741	392,981	NW_004451499	398,273	3.90	3.71	5.70
KQ079794	1,083,068	NW_004450885	1,098,196	3.98	3.80	5.78

Table 22: Run time of parallel scoring algorithms on Xeon24.

A	|A|	B	|B|	PP_Score	PP_Diagonal	PP_Strip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:02	0:00:04	0:00:01	52.7%	72.8%
XM_004403740	104,267	XM_012169741	103,004	0:00:02	0:00:04	0:00:01	46.3%	68.6%
AC006294	200,000	NW_009147196	200,000	0:00:10	0:00:18	0:00:04	56.5%	76.2%
AC006294	200,000	NW_004451499	398,273	0:00:20	0:00:31	0:00:09	56.2%	72.0%
KQ079741	392,981	NW_004451499	398,273	0:00:42	0:00:54	0:00:16	61.4%	69.8%
KQ079794	1,083,068	NW_004450885	1,098,196	0:04:33	0:05:36	0:01:57	57.2%	65.2%

Table 23: Speedup of parallel alignment algorithms on Xeon24.

A	|A|	B	|B|	Score/PP	Diagonal/PP	Strip/PP
human4	97,634	mouse4	94,647	10.17	6.50	18.90
XM_004403740	104,267	XM_012169741	103,004	10.63	8.45	18.05
AC006294	200,000	NW_009147196	200,000	10.03	6.03	19.94
AC006294	200,000	NW_004451499	398,273	9.98	7.90	18.32
KQ079741	392,981	NW_004451499	398,273	9.07	8.04	21.10
KQ079794	1,083,068	NW_004450885	1,098,196	11.36	9.56	22.22

Table 24: Run time of alignment algorithms, in hh:mm:ss, on Xeon24.

A	|A|	B	|B|	MM	MMDiagonal	MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:46	0:00:44	0:00:40	11.9%	7.6%
XM_004403740	104,267	XM_012169741	103,004	0:00:50	0:00:56	0:00:46	7.4%	16.7%
AC006294	200,000	NW_009147196	200,000	0:03:19	0:03:09	0:02:45	17.1%	12.5%
AC006294	200,000	NW_004451499	398,273	0:06:01	0:06:46	0:05:41	5.8%	16.2%
KQ079741	392,981	NW_004451499	398,273	0:12:56	0:13:30	0:11:09	13.8%	17.4%
KQ079794	1,083,068	NW_004450885	1,098,196	1:30:06	1:40:36	1:23:45	7.1%	16.8%

Table 25: Run time of parallel alignment algorithms on Xeon24.

A	|A|	B	|B|	PP_MM	PP_MMDiagonal	PP_MMStrip	Imp1	Imp2
human4	97,634	mouse4	94,647	0:00:05	0:00:22	0:00:03	28.9%	84.6%
XM_004403740	104,267	XM_012169741	103,004	0:00:06	0:00:25	0:00:05	29.4%	81.9%
AC006294	200,000	NW_009147196	200,000	0:00:23	0:00:58	0:00:14	41.3%	76.3%
AC006294	200,000	NW_004451499	398,273	0:00:43	0:01:48	0:00:24	43.6%	77.6%
KQ079741	392,981	NW_004451499	398,273	0:01:19	0:02:22	0:00:44	44.4%	69.0%
KQ079794	1,083,068	NW_004450885	1,098,196	0:09:49	0:15:04	0:05:10	47.3%	65.7%

Table 26: Speedup of alignment algorithms on Xeon24.

A	|A|	B	|B|	MM/PP	MMDiagonal/PP	MMStrip/PP
human4	97,634	mouse4	94,647	9.68	1.99	12.00
XM_004403740	104,267	XM_012169741	103,004	7.81	2.23	10.24
AC006294	200,000	NW_009147196	200,000	8.56	3.27	12.08
AC006294	200,000	NW_004451499	398,273	8.39	3.75	14.02
KQ079741	392,981	NW_004451499	398,273	9.79	5.70	15.17
KQ079794	1,083,068	NW_004450885	1,098,196	9.18	6.68	16.20

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 329-338 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Machine Learning Applied to GRBAS Voice Quality Assessment

Zheng Xie[*,1], Chaitanya Gadepalli[2], Farideh Jalalinajafabadi[3], Barry M.G. Cheetham[3], Jarrod J. Homer[4]

[1]*School of Engineering, University of Central Lancashire, PR1 2HE, UK*

[2]*Department of ENT, Salford Royal Hospital Foundation Trust, Salford, M6 8HD, UK*

[3]*School of Computer Science, University of Manchester, M13 9PL, UK*

[4]*Consultant in Head and Neck Surgery at the Manchester Royal Infirmary, Manchester University Hospitals Foundation Trust, UK*

ARTICLE INFO

Article history:
Received: 24 August, 2018
Accepted: 16 November, 2018
Online: 01 December, 2018

Keywords:
Voice quality assessment
GRBAS, Consistency measures
Cohen Kappa, Fleiss Kappa
Intra-class correlation
Feature detection
Machine learning

ABSTRACT

Voice problems are routinely assessed in hospital voice clinics by speech and language therapists (SLTs) who are highly skilled in making audio-perceptual evaluations of voice quality. The evaluations are often presented numerically in the form of five-dimensional 'GRBAS' scores. Computerised voice quality assessment may be carried out using digital signal processing (DSP) techniques which process recorded segments of a patient's voice to measure certain acoustic features such as periodicity, jitter and shimmer. However, these acoustic features are often not obviously related to GRBAS scores that are widely recognised and understood by clinicians. This paper investigates the use of machine learning (ML) for mapping acoustic feature measurements to more familiar GRBAS scores. The training of the ML algorithms requires accurate and reliable GRBAS assessments of a representative set of voice recordings, together with corresponding acoustic feature measurements. Such 'reference' GRBAS assessments were obtained in this work by engaging a number of highly trained SLTs as raters to independently score each voice recording. Clearly, the consistency of the scoring is of interest, and it is possible to measure this consistency and take it into account when computing the reference scores, thus increasing their accuracy and reliability. The properties of well known techniques for the measurement of consistency, such as intra-class correlation (ICC) and the Cohen and Fleiss Kappas, are studied and compared for the purposes of this paper. Two basic ML techniques, i.e. K-nearest neighbour regression and multiple linear regression were evaluated for producing the required GRBAS scores by computer. Both were found to produce reasonable accuracy according to a repeated cross-validation test.

1. Introduction

Voice problems are a common reason for referrals by primary practices to ear, nose and throat (ENT) departments and voice clinics in hospitals. Such problems may result from voice-strain due to speaking or singing excessively or too loudly, vocal cord inflammation, side-effects of inhaled steroids as used to treat asthma, infections, trauma, neoplasm, neurological disease and many other causes. This paper is an extension of work on voice quality assessment originally presented in the 10th CISP-BMEI, conference in Shanghai [1]. Speech and language therapists (SLTs) are commonly required to assess the nature of voice quality impairment in patients, by audio-perception. This requires the SLT, trained as a voice quality expert, to listen to and assess the patient's voice while it reproduces, or tries to reproduce, certain standardized vocal maneuvers. In Europe, voice quality assessments are often made according to the perception of five properties of the voice as proposed by Hirano [2]. The five properties are referred to by the acronym 'GRBAS' which stands

[*]Corresponding Author: Zheng Xie, Email: zxie2@uclan.ac.uk

for 'grade', 'roughness', 'breathiness', 'asthenia' and 'strain'. Each GRBAS property is rated, or scored by assigning an integer 0, 1, 2 or 3. A score of 0 signifies no perceived loss of quality in that property, 1 signifies mild loss of quality, 2 signifies moderate loss and 3 signifies severe loss. The scoring may be considered categorical or ordinal. With categorical scoring the integers 0, 1, 2 and 3 are considered as labels. With ordinal scoring, the integers are considered as being numerical with magnitudes indicating the severity of the perceived quality loss.

Grade (G) quantifies the overall perception of voice quality which will be adversely affected by any abnormality. Roughness (R) measures the perceived effect of uncontrolled irregular variations in the fundamental-frequency and amplitude of vowel segments which should be strongly periodic. Breathiness (B) quantifies the level of sound that arises from turbulent air-flow passing through vocal cords when they are not completely closed. Asthenia (A) measures the perception of weakness or lack of energy in the voice. Strain (S) gives a measure of undue effort needed to produce speech when the speaker is unable to employ the vocal cords normally because of some impairment.

Voice quality evaluation by audio-perception is time-consuming and expensive in its reliance on highly trained SLTs [3]. Also, inter-rater inconsistencies must be anticipated, and have been observed [4] in the audio-perceptual scoring of groups of patients, or their recorded voices, by different clinicians. Intra-rater inconsistencies have also been observed when the same clinician re-assesses the same voice recordings on a subsequent occasion. A lack of consistency in GRBAS assessments can adversely affect the appropriateness of treatment offered to patients, and the monitoring of its effect. A computerised approach to GRBAS assessment could eliminate these inconsistencies.

According to Webb et al. [5], GRBAS is simpler and more reliable than many other perceptual voice evaluation scales, such as Vocal Profile Analysis (VPA) [6] and the 'Buffalo Voice Profile' (BVP) [7], scheme. The 'Consensus Auditory-Perceptual Evaluation of Voice' (CAPE-V) approach, as widely used in North America [8], allows perceptual assessments of overall severity, roughness, breathiness, strain, pitch and loudness to be expressed as percentage scores. It is argued [8] that, compared with GRBAS, the CAPE-V scale better measures the quality of the voice and other aphonic characteristics. Also, CAPE-V assessments are made on a more refined scale. However, GRBAS is widely adopted [9] by practising UK voice clinicians as a basic standard.

No definitive solutions yet exist for performing GRBAS assessments by computer. Some approaches succeeded in establishing reasonable correlation between computerised measurements of acoustic voice features and GRBAS scores, but have not progressed to prototype systems [12]. Viable systems have been proposed, for example [13], but problems of training the required machine learning algorithms remain to be solved. The 'Multi-Dimensional Voice Program' (MDVP) and 'Analysis of Dysphonia in Speech and Voice' (ADSV) are commercial software packages [10] providing a wide range of facilities for acoustic feature analysis. Additionally, ADSV gives an overall assessment of voice dysphonia referred to as the Cepstral/spectral Index of Dysphonia (CSID) [11]. This is calculated from a multiple regression based on the correlation of results from ADSV analyses with CAPE-V perceptual analyses by trained scorers. The CAPE-V overall measure of dysphonia is closely related to the 'Grade' component of GRBAS, therefore the CSID approach offers a methodology and partial solution to the GRBAS prediction problem. However the commercial nature of the CSID software makes it difficult to study and build on this methodology. Therefore, this paper considers how the results of a GRBAS scoring exercise may be used to produce a set of reference scores for training machine learning algorithms for computerised GRBAS assessment.

For the purposes of this research, a scoring exercise was carried out with the participation of five expert SLT raters, all of whom were trained and experienced in GRBAS scoring and had been working in university teaching hospitals for more than five years. A database of voice recordings from 64 patients was accumulated over a period of about three months by randomly sampling the attendance at a typical voice clinic. This database was augmented by recordings obtained from 38 other volunteers.

The recordings were made in a quiet studio at the Manchester Royal Infirmary (MRI) Hospital. Ethical approval was given by the National Ethics Research committee (09/H1010/65). The KayPentax 4500 CSL ® system and a Shure SM48 ® microphone were used to record the voices with a microphone set at 45 degrees at a distance of 4 cm. The recordings were of sustained vowel sounds and segments of connected speech.

To obtain the required GRBAS scores for each of the subjects (patients and other volunteers), the GRBAS properties of the recordings were assessed independently by the five expert SLT raters with the aid of a 'GRBAS Presentation and Scoring Package (GPSP)' [14]. This application plays out the recorded sound and prompts the rater to enter GRBAS scores. Raters used Sennheiser HD205 ® head-phones to listen to the recorded voice samples. The voice samples are presented in randomised order with a percentage (about 20 %) of randomly selected recordings repeated without warning, as a means of allowing the self-consistency of each rater to be estimated.

Different statistical methods were then employed to measure the intra-rater consistency (self-consistency) and inter-rater consistency of the scoring. Some details of these methods are presented in the next section. The derivation of 'reference' GRBAS scores from the audio-perceptual rater scores is then considered for the purpose of training ML algorithms for computerised GRBAS scoring. The derivation takes into account the inter-rater and intra-rater consistencies of each rater,

Voice quality assessment may be computerised using digital signal processing (DSP) techniques which analyse recorded segments of voice to quantify universally recognised acoustic features such as fundamental frequency, shimmer, jitter and harmonic-to-noise ratio [14]. Such acoustic features are not obviously related to the GRBAS measurements that are widely recognised and understood by clinicians. We therefore investigated the use of machine learning (ML) for mapping these feature measurements to the more familiar GRBAS assessments. Our approach was to derive 'reference scores' for a database of voice recordings from the scores given by expert SLT raters. The reference scores are then used to train a machine-learning

algorithm to predict GRBAS scores from the acoustic feature measurements resulting from the DSP analysis. The effectiveness of these techniques for computerised GRBAS scoring is investigated in Section 13 of this paper.

2. Measurement of Consistency

The properties of a number of well-known statistical methods for measuring rater consistency were considered for this research. The degree of consistency between two raters when they numerically appraise the same phenomena may be measured by a form of correlation. Perhaps the best known form of correlation is Pearson Correlation [15]. However, this measure takes into account only variations about the individual mean score for each rater [16]. Therefore a rater with consistently larger scores than those of another rater can appear perfectly correlated and therefore consistent with that other rater. Pearson correlation has been termed a measure of 'reliability' [17] rather than consistency. It is applicable only to ordinal appraisals, and is generally inappropriate for measuring consistency between or among raters [9] where consistency implies agreement. The notion of consistency between two raters can be extended to self-consistency between repeated appraisals of the same phenomena by the same rater (test-retest consistency), and to multi-rater consistency among more than two raters.

An alternative form of correlation is given by the 'intra-class correlation' coefficient (ICC) [18] and this may be used successfully as a measure of consistency for rater-pairs. It is also suitable for intra-rater (test-retest) and multi-rater consistency. The scoring must be ordinal. ICC is based on the differences that exist between the scores of each rater and a 'pooled' arithmetic mean score that is computed over all the scores given by all the raters. Therefore ICC eliminates the disadvantage of Pearson Correlation that it takes into account only variations about the individual mean score for each rater.

The 'proportion of agreement' (P_o), for two raters, is a simple measure of their consistency. It is derived by counting the number of times that the scores agree and dividing by the number, N, of subjects. P_o will always be a number between 0 (signifying no agreement at all) and 1 (for complete agreement). It is primarily for categorical scoring but may also be applied to ordinal scoring where the numerical scores are considered as labels. For ordinal scoring, P_o does not reflect the magnitudes of any differences, and in both cases, P_o is biased by the possibility of agreement by chance. The expectation of P_o will not be zero for purely random scores because some of the scores will inevitably turn out to be equal by chance. With Q different categories or scores evenly distributed over the Q possibilities, the probability of scores being equal by chance would be 1/Q. Therefore, the expectation of P_o would be $1/Q$ rather than zero for purely random scoring. With $Q = 4$, this expectation would be a bias of 0.25 in the value of P_o. The bias could be even greater with an uneven spread of scores by either rater. The bias may give a false impression of some consistency when there is none, as could occur when the scores are randomly generated without reference to the subjects at all.

The Cohen Kappa is a well known consistency measure originally defined [19] for categorical scoring by two raters. It was later generalised to the weighted Cohen Kappa [20] which is applicable to ordinal (numerical) scoring with the magnitudes of any disagreements between scores taken into account. The Fleiss Kappa [21] is a slightly different measure of consistency for categorical scoring that may be applied to two or more raters. The significance of Kappa and ICC measurements is often summarised by descriptions [22, 23] that are reproduced in Tables 1 and 2. A corresponding table for the Pearson correlation coefficient may be found in the literature [24].

Table 1: Significance of Kappa Values

Kappa	Consistency
1.0	Perfect
0.8 – 1.0	Almost perfect
0.6 - 0.8	Substantial
0.4 - 0.6	Moderate
0.2 - 0.4	Fair
0 - 0.2	Slight
< 0	Less than chance

Table 2: Significance of ICC Values

ICC	Consistency
0.75 – 1.0	Excellent
0.4 - 0.75	Fair
< 0.4	Poor

3. The Cohen Kappa

The original Cohen Kappa [19] for two raters, A and B say, was defined as follows:

$$Kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the proportion of agreement, as defined above, and P_e is an estimate of the probability of agreement by chance when scores by two raters are random (unrelated to the patients) but distributed across the range of possible scores identically to the actual scores of raters A and B. The estimate P_e is computed as the proportion of subject pairs (i, j) for which the score given by rater A to subject i is equal to the score given by rater B to subject j. This is an estimate of the probability that a randomly chosen ordered pair of subjects (i, j) will have equal scores.

This measure of consistency [19] is primarily for categorical scoring, though it can be applied to ordinal scores considered as labels. In this case, any difference between two scores will be considered equally significant, regardless of its numerical value. Therefore, it will only be of interest whether the scores, or classifications, are the same or different.

The weighted Cohen Kappa [20] measures the consistency of ordinal scoring where numerical differences between scores are considered important. It calculates a 'cost' for each actual disagreement and also for each expected 'by chance' disagreement. The cost is weighted according to the magnitude of the difference between the unequal scores. To achieve this, equation (1) is re-expressed by equation (2):

$$Kappa = 1 - \frac{1 - P_o}{1 - P_e} = 1 - \frac{D_o}{D_e} \tag{2}$$

where $D_o = 1 - P_o$ is the proportion of actual scores that are not equal and is considered to be the accumulated cost of the disagreements. The quantity $D_e = 1 - P_e$ is now considered to be the accumulated cost of disagreements expected to occur 'by chance' with random scoring distributed identically to the actual scores. Weighting is introduced by expressing D_o and D_e in the form of equations (3) and (4):

$$D_o = \frac{1}{N}\sum_{i=1}^{N} C(A(i), B(i)) \tag{3}$$

$$D_e = \frac{1}{N^2}\sum_{i=1}^{Q}\sum_{j=1}^{Q} A_i B_j C(\alpha(i), \alpha(j)) \tag{4}$$

In equations (3) and (4), $C(a,b)$ is the cost of any difference between scores (or categories) a and b. In equation (4), A_i denotes the number of subjects that rater A scores as $\alpha(i)$ and B_j denotes the number of subjects that rater B scores as $\alpha(j)$. Q is the number of possible scores or scoring categories and these are denoted by $\alpha(1)$, $\alpha(2)$… $\alpha(Q)$. If the cost-function C is defined by equation (5):

$$C(a,b) = \begin{cases} 1 : a \neq b \\ 0 : a = b \end{cases} \tag{5}$$

the weighted Cohen Kappa [20] becomes identical to the original Cohen Kappa [19] also referred to as the unweighted Cohen Kappa (*UwCK)*. If C is defined by equation (6),

$$C(a,b) = |a - b| \tag{6}$$

we obtain the 'linearly weighted Cohen Kappa' (*LwCK*), and defining C by equation (7) produces the 'quadratically weighted Cohen Kappa'(*QwCK*).

$$C(a,b) = (a - b)^2 \tag{7}$$

There are other cost-functions with interesting properties, but the three mentioned above are of special interest. For GRBAS scoring, there are Q = 4 possible scores which are $\alpha(1)$=0, $\alpha(2)$=1, $\alpha(3)$=2 and $\alpha(4)$=3.

Equation (4) may be re-expressed as equation (8):

$$D_e = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i), B(j)) \tag{8}$$

Therefore, from equations (2), (3) and (8), we obtain equation (9) which is a general formula for all 2-rater (pair-wise) forms of Cohen Kappa:

$$Kappa = 1 - \frac{(1/N)\sum_{i=1}^{N} C(A(i), B(i))}{(1/N^2)\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i), B(j))} \tag{9}$$

The original and weighted Cohen Kappa [19, 20] are applicable when there are two individual raters, A and B say, who both score all the N subjects. The raters are 'fixed' in the sense that rater A is always the same clinician who sees all the subjects; and similarly for rater B. Therefore the individualities and prejudices of each rater can be taken into account when computing P_e, the probability of agreement by chance. For example, if one rater tends to give scores that are consistently higher than those of the other rater, this bias will be reflected in the value of Cohen Kappa obtained.

4. Other Versions of Kappa

The Fleiss Kappa [21] measures the consistency of two or more categorical raters, and can therefore be a 'multi-rater' consistency measure. Further, the raters are not assumed to be 'fixed' since each subject may be scored by a different pair or set of raters. Therefore, it is no longer appropriate to take into account the different scoring preferences of each rater. If the Fleiss Kappa is used for a pair of fixed raters as for the Cohen Kappas, slightly different measurements of consistency will be obtained.

Assuming that there are n raters and Q scoring categories, Fleiss [21] calculates the proportion p_j of the N subjects that are assigned by raters to category j, as follows:

$$p_j = \frac{1}{N \times n}\sum_{i=1}^{N} n_{ij} \tag{10}$$

for j = 1, 2, …, Q, where n_{ij} is the number of raters who score subject i as being in category j. The proportion, P_i, of rater-pairs who agree in their scoring of subject i is given by:

$$P_i = \frac{1}{L}\sum_{j=1}^{Q} n_{ij} \times (n_{ij} - 1)/2 \tag{11}$$

where L is the number of different rater-pairs that are possible, i.e. $L = n(n\text{-}1)/2$. The proportion of rater-pairs that agree in their assignments, taking into account all raters and all subjects, is now:

$$P_o = \frac{1}{N}\sum_{i=1}^{N} P_i \tag{12}$$

Fleiss [21] then estimates the probability of agreement 'by chance' as:

$$P_e = \sum_{j=1}^{Q} p_j^{\,2} \tag{13}$$

Substituting from equations (12) and (13) into the Kappa equation (1) gives the Fleiss Kappa [21] which may be evaluated for two or more raters not assumed to be 'fixed' raters. The resulting equation does not generalise the original Cohen Kappa because equation (13) does not take any account of how the scores by each individual rater are distributed. P_e is now dependent only on the overall distribution of scores taking all raters together. Agreement by chance is therefore redefined for the Fleiss Kappa.

The original Cohen Kappa may be truly generalised [27] to measure the multi-rater consistency of categorical scoring by a group of n 'fixed' raters, where $n \geq 2$. Light [28] and Hubert [29] published different versions for categorical scoring, and Conger [30] extended the version by Light [28] to more than three raters. The generalisation by Hubert [29] redefines D_o and D_e to include all possible rater-pairs as in equation (14):

$$D_o = \frac{1}{L}\sum_{r=1}^{n}\sum_{s=r+1}^{n} D_o(r,s) \quad D_e = \frac{1}{L}\sum_{r=1}^{n}\sum_{s=r+1}^{n} D_e(r,s) \tag{14}$$

where the expression for $D_o(r,s)$ generalises equation (3) and the expression for $D_e(r,s)$ generalises equation (8) to become the cost of actual disagreement and the expected cost of by chance disagreement between raters r and s. Denoting by $A(i, r)$ the score given by rater r to subject i, we obtain equations (15) and (16):

$$D_o(r,s) = \frac{1}{N}\sum_{i=1}^{N} C(A(i,r), A(i,s)) \tag{15}$$

$$D_e(r,s) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} C(A(i,r), A(j,s)) \tag{16}$$

where equation (5) defines the cost-function C. Substituting for D_o and D_e from equation (14) into equation (2), with $D_o(r,s)$ and $D_e(r,s)$ defined by equations (15) and (16) gives a formula for the multi-rater Cohen Kappa that is functionally equivalent to that published by Hubert [29]. With C defined by equation (5) it remains unweighted.

The generalisation by Light [28] is different from the Hubert version when $n > 2$. It is given by equation (17):

$$UwCK = 1 - \frac{1}{L}\sum_{r=1}^{n}\sum_{s=r+1}^{n}\frac{D_o(r,s)}{D_e(r,s)} \tag{17}$$

Although both generalisations were defined for categorical scoring, they may now be further generalised to weighted ordinal scoring simply by redefining the cost-function C, for example by equation (6) for linear weighting or equation (7) for quadratic weighting. With $n = 2$, both generalisations are identical to the original [19] or weighted [20] Cohen Kappa.

5. Weighted Fleiss Kappa

As explained in [31], the original Fleiss Kappa [21] is given by equation (18) when the cost-function C is as in equation (5).

$$FK = 1 - \frac{\frac{1}{NL}\sum_{i=1}^{N}\sum_{r=1}^{n}\sum_{s=r+1}^{n} C(A(i,r), A(i,s))}{\frac{1}{(Nn)^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{r=1}^{n}\sum_{s=1}^{n} C(A(i,r), A(j,s))} \tag{18}$$

The Fleiss Kappa may be generalised to a weighted version for ordinal scoring by redefining cost-function C as for the multi-rater Cohen Kappa. In all cases, the unweighted or weighted Fleiss Kappa is applicable to measuring the consistency of any number of raters including two.

6. Intra-Class Correlation Coefficient (*ICC*)

In its original form [25], *ICC* is defined for n raters as follows:

$$ICC = \frac{(1/L)\sum_{i=1}^{N}\sum_{r=1}^{n}\sum_{s=r+1}^{n}(A(i,r)-m)(A(i,s)-m)}{(1/n)\sum_{i=1}^{N}\sum_{r=1}^{n}(A(i,r)-m)^2} \tag{19}$$

$$\text{where} \quad m = \frac{1}{nN}\sum_{i=1}^{N}\sum_{r=1}^{n} A(i,r) \tag{20}$$

Other versions of *ICC* have also been proposed [26]. It is known [26] that, for two raters, *ICC* will be close to quadratically weighted Cohen Kappa (*QwCK*) when the individual mean score for each rater is approximately the same. This property is observed [31] also for multi-rater versions of *ICC* and *QwCK*. More interestingly, it has been shown [31] that *ICC* is always exactly equal to quadratically weighted Fleiss Kappa (*QwFK*) regardless of the number of raters and their individual mean scores.

7. Intra-rater Consistency

For the GRBAS rating exercise referred to in Section 1, intra-rater (test-retest) scoring differences were generally small due to the experience and high expertise of the SLT raters. There were some differences of 1, very occasional differences of 2, and no greater differences. The test-retest consistency for the five GRBAS components was measured for all five raters, by unweighted, linearly weighted and quadratically weighted Cohen Kappa (*UwCK*, *LwCK* and *QwCK*) and *ICC*. By averaging *UwCK*, *LwCK* and pair-wise *ICC* measurements over the five GRBAS components we obtained Table 3 which gives three overall measurements of the test-retest consistency of each rater. *QwCK* gave a close approximation to *ICC,* and is not shown in the table. *QwFK,* also not shown, was indistinguishable from *ICC*. For all forms of Kappa, the P_o and P_e terms were averaged separately. Similarly, the *ICC* numerators and denominators were averaged separately.

With UwCK, any difference in scores incurs the same cost regardless of its magnitude. Small differences cost the same as large differences. This makes *UwCK* pessimistic for highly consistent raters where most test-retest discrepancies are small. Therefore, the averaged *UwCK* consistency measurements in Table 3 are pessimistic for our rating exercise.

With *QwCK*, the largest differences in scores incur very high cost due to the quadratic weighting. With ICC, the costs are similar. These high costs are important even when there are few or no large scoring differences because they strongly affect the costs of differences expected to incur 'by chance'. These high 'by chance' costs make both *QwCK* and *ICC* optimistic, when compared with *LwCK,* for highly consistent rating with a fairly even distribution of scores. We therefore concluded that *LwCK* gives the most indicative measure of test-retest consistency for the rating exercise referred to in this paper. A different set of scores may have led to a different conclusion. In Table 3, it may be seen that the self-consistency of raters 1 to 4, as measured by *LwCK*, was considered 'substantial' according to Table 2. The self-consistency of rater 5 was considered 'moderate'. Conclusions can therefore be drawn about the self-consistency of each rater and how this may be expected to vary from rater to rater.

Table 3: Intra-Rater Consistency Averaged over all GRBAS Components

Rater	*UwCK*	*LwCK*	*ICC*	Consistncy (*LwCK*)	Consistncy (*ICC*)
1	0.72	0.77	0.84	Substantial	Excellent
2	0.65	0.76	0.85	Substantial	Excellent
3	0.53	0.64	0.75	Substantial	Excellent
4	0.68	0.73	0.77	Substantial	Excellent
5	0.44	0.60	0.74	Moderate	Fair

Table 4: Intra-rater Consistency Averaged over all 5 raters

Comp-onent	UwCK	LwCK	ICC	Consistncy (LwCK)	Consistncy (ICC)
G	0.64	0.77	0.87	Substantial	Excellent
R	0.57	0.67	0.76	Substantial	Excellent
B	0.55	0.66	0.76	Substantial	Excellent
A	0.68	0.73	0.80	Substantial	Excellent
S	0.59	0.68	0.76	Substantial	Excellent

Table 4 shows *UwCK*, *LwCK* and *ICC* intra-rater consistency measurements for G, R, B, A, and S, averaged over all five raters. According to all the measurements, it appears that test-retest consistency with R, B and S is more difficult to achieve than with G and A.

8. Inter-rater Consistency

Measurements of inter-rater consistency between pairs of raters for any GRBAS component may be obtained using the same forms of Kappa and ICC as were used for intra-rater consistency. Our rating exercise had a group of five raters, therefore ten possible pairs. This means that there are ten pair-wise measurements of inter-rater consistency for each GRBAS component. To reduce the number of measurements, it is convenient to define an 'individualised' inter-rater measurement for each rater. For each GRBAS component, this individualised measurement quantifies the consistency of the rater with the other raters in the group. It is computed for each rater by averaging all the pair-wise inter-rater assessments which involve that rater. Thus an individualised measure of inter-rater consistency is obtained for G, R, B, A and S for each rater. With five raters, the 25 measurements can be reduced to five by averaging the individualised G, R, B, A and S measurements to obtain a single average measure for each rater.

The *UwCK*, *LwCK* and *ICC* individualised inter-rater measurements, averaged over all GRBAS components, are shown in Table 5 for raters 1 to 5. For all raters, the average consistency is 'moderate' according to *LwCK* and 'fair' according to *ICC*. Raters 1, 4 and 5 have almost the same inter-rater consistency, rater 2 has slightly lower consistency and rater 2 is the least consistent when compared with the other raters.

Table 5: Individualised Inter-rater Consistency averaged over all GRBAS Components

Rater	*UwCK*	*LwCK*	*ICC*	Consistncy (*LwCK*)	Consistncy (*ICC*)
1	0.47	0.59	0.70	Moderate	Fair
2	0.40	0.52	0.60	Moderate	Fair
3	0.45	0.57	0.67	Moderate	Fair
4	0.48	0.60	0.71	Moderate	Fair
5	0.47	0.59	0.70	Moderate	Fair

9. Multi-rater Consistency

The multi-rater consistency according to the unweighted Fleiss Kappa (FK), the generalised Cohen Kappa (with linear weighting) and *ICC,* computed for the group of five raters, are shown in Table 6 for each GRBAS component. The values of *UwCK* were indistinguishable from FK to the precision shown in the table. Similarly for the values of *QwCK* and *ICC*. Quadratically weighted *FK,* also not shown, would be exactly equal to *ICC*.

Table 6: Multi-rater Consistency by Fleiss Kappa, Cohen Kappa and *ICC*

	FK	*LwCK*	*ICC*	Consistncy (LwCK)	Consistncy (*ICC*)
G	0.56	0.71	0.83	Substantial	Excellent
R	0.44	0.57	0.68	Moderate	Fair
B	0.43	0.58	0.71	Moderate	Fair
A	0.38	0.46	0.55	Moderate	Fair
S	0.44	0.54	0.65	Moderate	Fair

In contrast to Table 5 which allows us to compare the overall consistency of raters, Table 6 allows us to compare the difficulty of achieving group consistency for each GRBAS component. It is clear that some GRBAS components are more difficult to score consistently than others. According to *ICC*, group consistency is 'excellent' for Grade and 'fair' for R, B, A and S. *LwCK* gives 'substantial' for Grade and 'fair' for the others. The *FK* and *UwCK* measurements are more pessimistic due to their assumption that the scoring is categorical. According to all measurements of multi-rater consistency, the consistency is highest for highest, followed by Breathiness, Roughness, Strain and Asthenia.

It should be mentioned that the classifications given by Tables 1 and 2 serve only as a rough guide to interpreting the values of Kappa and ICC obtained. However, they are widely used despite the fact that it seems inappropriate to use Table 2 for quadratically weighted Kappa in view of its closeness to ICC. In particular, the category 'Fair' in Tables 2 and 3 refers to quite different ranges which may be misleading if Table 2 were used for QwCK. Therefore, it is appropriate to refer to Table 3 for both ICC and quadratically weighted Kappa.

10. Reference GRBAS Scores

The feasibility of performing automatic GRBAS scoring by computer was investigated by training machine learning (ML) algorithms for mapping acoustic feature measurements to the familiar GRBAS scale. For the training, a set of accurate and reliable GRBAS scores was required for each of the *N* subjects in our database. We refer to these as 'reference' GRBAS scores. A technique for deriving these reference scores from the scores of a group of audio-perceptual raters, such as that described in Section 1, was therefore devised. The measurements of inter-rater and intra-rater consistency, obtained as described above, is taken into account as a means of optimising the accuracy and reliability of the reference scores.

Given the 'Grade' scores *A(i, r)* for subject *i,* with rater-index *r* in the range 1 to 5, we first computed weighted average pair-wise scores $G_{rs}(i)$ by equation (21), for all possible rater-pairs *(r,s)*. The weighting is by the *LwCK* intra-rater consistency measurements in Table 3 referred to now as w_1, w_2, w_3, w_4, w_5 for raters 1 to 5 respectively.

$$G_{rs}(i) = \frac{w_r A(i,r) + w_s A(i,s)}{w_r + w_s} \tag{21}$$

The 'Grade' reference score for subject *i* is then obtained as a weighted average of the $G_{rs}(i)$ values over all possible rater-pairs, i.e.:

$$G_{ref}(i) = \frac{1}{L}\sum_{r=1}^{n}\sum_{s=r+1}^{n} w(r,s)G_{rs}(i) \quad (22)$$

where $L = n(n-1)/2$ with $n=5$. The weights $w(r,s)$ are the pair-wise inter-rater *LwCK* measurements for Grade. This procedure is performed for all subjects for Grade, and then repeated for the other GRBAS dimensions. The weighting de-emphasises scores from less self-consistent raters in favour of more self-consistent ones. It also de-emphasises the scores from raters who are less consistent with other raters.

11. Voice Quality Assessment by Computer

Considerable published research, including [12] and [13], has not yet established a definitive methodology for GRBAS assessment by computer. An overall CAPE-V assessment of dysphonia, CSID [11], available commercially, is strongly related to 'Grade', but it does not independently assess the other GRBAS and CAPE-V components [8]. Computerised voice quality assessment may be carried out using digital signal processing (DSP) to analyse segments of voice to produce mathematical functions such as the autocorrelation function, fast Fourier Transform and cepstrum. From such functions, acoustic features such as the aperiodicity index (*API*), fundamental frequency (F_0), harmonic-to-noise ratio (*HNR*), jitter, shimmer, cepstral peak prominence (*CPP*), low-to-high spectral ratio (*LH*) and others may be derived. However, these features are not obviously related to GRBAS assessments of voice quality.

Perceived voice quality is strongly dependent on the short term periodicity of the vowels and the nature of the fluctuations in this periodicity. To measure short-term periodicity, and how this varies over a spoken vowel, speech must be segmented into frames. The degree of periodicity of each of these frames may expressed as an aperiodicity index (API) which is equal to 1 - p where p is the peak value of a suitable form of autocorrelation function. An API of zero indicates exact periodicity and its value increases towards 1 with increasing aperiodicity. The *API* is increased by additive noise due to 'breathiness', fundamental frequency or amplitude variation due to 'roughness' in the operation of the vocal cords, and other acoustic features.

A sustained vowel without obvious impairment will generally have strong short-term periodicity for the duration of the segment, though the fundamental frequency (F_0) and loudness may vary due to natural characteristics of the voice and controlled intonation. By monitoring how the degree of short term periodicity changes over a passage of natural connected speech, vowels may be differentiated from consonants, thus allowing the acoustic feature measurements to concentrate on the vowels.

Jitter is rapid and uncontrolled variation of F_0 and shimmer is rapid and uncontrolled variation of amplitude. Both these acoustic features can be indicative of roughness in GRBAS assessments. They will affect grade also. There are many ways of defining jitter and shimmer as provided by the Praat software package [32]. The HNR may be derived from the autocorrelation function and can be indicative of breathiness in GRBAS assessments since the 'noise' is often due to turbulent airflow. Low-to-high spectral ratio (LH) measurements are made by calculating and comparing, in the frequency-domain, the energy below and above a certain cut-off frequency, such as 1.5 kHz or 4.0 kHz. The required filtering may be achieved either by digital filters or an FFT. A high value of LH with cut-off frequency 1.5 kHz can be indicative of asthenia [36] and strain [37] due to imperfectly functioning vocal cords damping the spectral energy of formants above 1.5 kHz. LH measurements with a cut-off frequency of 4.0 kHz are useful for detecting breathiness and voicing since the spectral energy of voiced speech (vowels) is mostly below 4.0 kHz. CPP is widely used as an alternative to API and HNR as a means of assessing the degree of short term periodicity.

As in [34], well known DSP techniques were employed [14, 35] to recognise vowels and measure the acoustic features mentioned above, and several others. Frame-to-frame variations in these features over time were also measured. Published DSP algorithms and commercial and academic computer software are available for making these measurements from digitised voice recordings [32, 33]. Twenty acoustic features were identified by Jalalinajafabadi [14] as being relevant to GRBAS scores. They were measured by a combination of DSP algorithms specially written in MATLAB and commercial software provided by MDVP and ADSV [10, 11]. For the MATLAB algorithms, the speech recordings were sampled at Fs = 44.1 kHz, and divided into sequences of 75% overlapping 23.22 ms frames of 1024 samples. MDVP and ADSV use a slightly different sampling rate and framing. Many of the features were strongly correlated and their usefulness was far from uniform. Therefore, some experiments with feature selection were performed. The usefulness of each possible sub-set of features for predicting each GRBAS component was estimated by a combination of correlation measurements, to reduce the dimensionality of the task, and then a form of direct search. The use of Principle Component Analysis' (PCA) would have reduced the computation, but this was not a critical factor.

Section 14 will evaluate the performance of MLR and KNNR (with and without feature selection) and perceptual analysis against the 'reference GRBAS scores'.

12. Machine Learning Algorithms

We analysed the recordings of sustained vowels obtained from the $N = 102$ subjects mentioned in Section 1. For each recording, acoustic feature measurements were obtained as explained in Section 11. A total of $m = 20$ feature measurements were obtained as detailed in [14]. An $N \times m$ matrix X of feature measurements was defined for each of the five GRBAS components. These matrices became the input to the machine learning (ML) algorithm along with the $N \times 1$ vector $\underline{Y}$ of reference GRBAS scores derived as explained in Section 10. The ML algorithm was designed to learn to predict, as closely as possible, the reference GRBAS scores supplied for each subject. The prediction must be made from the information provided by the m acoustic feature measurements supplied for each voice segment. Two simple ML approaches were compared [14, 35]: K-nearest neighbour regression (KNNR) and multiple linear regression (MLR).

With KNNR, the ML information consists of a matrix X and vector $\underline{Y}$ for each GRBAS component. Supplying the ML algorithm with these arrays is all that is required of the training process. K is an integer that defines the way the KNNR approach

predicts a score for a new subject from measurements of its *m* acoustic features. The prediction is based on the known scores for *K* other subjects chosen according to the 'distance' of their measured acoustic features from those of the new subject. The concept of distance can be defined in various ways such as the Euclidean distance which we adopted. The distance between the new subject and each of the *N* database subjects is calculated. Then *K* subjects are selected as being those that are nearest to the new subject according to their feature measurements. A simple form of KNNR takes the arithmetic mean of the scores of the *K* nearest neighbours as the result. A preferred alternative form takes a weighted average where the reference scores are weighted according to the proximity of the reference subject to the new subject.

A choice of K must be made, and this may be different for each GRBAS component. The optimal value of K will depend on the number, *N*, of subjects, the distribution of their scores and the number of acoustic features being taken into account. K is often set equal the square root of N, though investigations can reveal more appropriate values. In this work, Jalalinajafabadi [14] plotted the prediction error against K to obtain a suitable value of K for each GRBAS component. This was done after selecting the most appropriate set of acoustic features for each GRBAS component. The values of K producing the lowest prediction errors were K=6 for grade, K=10 for roughness, K=5 for breathiness and K = 8 for strain and asthenia.

The Multiple Linear Regression (MLR) approach computes, for each GRBAS component, a vector $\underline{\beta}$ of *K* regression coefficients such that

$$\underline{Y} = X.\underline{\beta} + \underline{\varepsilon} \tag{23}$$

where the error-vector $\underline{\varepsilon}$ is minimised in mean square value over all possible choices of $\underline{\beta}$ of dimension *K*. It may be shown [14] that the required vector $\underline{\beta}$ is given by:

$$\underline{\beta} = X^{\#}.\underline{Y} \tag{24}$$

where $X^{\#}$ is the pseudo-inverse of the non-square matrix X. For a subject whose *m* feature measurements $\underline{x}$ have been obtained, the equation:

$$y = \underline{x}^{T}.\underline{\beta} \tag{25}$$

produces a scoring estimate y. This will be close to *Y(i)* for each subject *i* in our database, and may be expected to produce reasonable GRBAS scores for an unknown subject.

13. Testing and Evaluation

The application developed by Jalalinajafabadi [14] made *m* = 20 voice feature measurements per subject. Feature selection was applied to identify which subset of these *m* features gave the best result for each GRBAS dimension. It was generally found that, compared with including all 20 feature measurements, better results were obtained with smaller subsets tailored to the GRBAS dimensions. Several computational methods for feature selection were compared [14] in terms of their effectiveness and computational requirements. The results presented here were obtained using a combination of correlation tables (between feature measurements and GRBAS components) and exhaustive search. The best feature subsets for G, R, B, A and S are generally different, since different feature measurements highlight different aspects of the voice. It was found beneficial to normalise the feature measurements to avoid large magnitudes dominating the prediction process, especially for KNNR.

To evaluate the KNNR and MLR algorithms for mapping acoustic feature measurements to GRBAS scores, 80 subjects were randomly selected for training purposes from the 102 available subjects. The remaining 22 subjects were set aside to be used for testing the mapping algorithms once they had been trained. Twenty 'trials' were performed by repeating the training and testing, each time with a different randomisation. The same testing approach was used for both KNNR and MLR. The trained mapping algorithm was used to predict GRBAS scores for the 22 testing subjects from the corresponding acoustic feature measurements. The GRBAS scores thus obtained were compared with the known reference scores. For each trial, a value of 'root mean squared error' (RMSE) was computed for each GRBAS component over the 22 testing subjects. These RMSE values were then averaged over the 20 trials. An RMSE of 100% would correspond to an RMS error of 1 in the GRBAS scoring where the averaging is over all 22 testing subjects and all 20 trials.

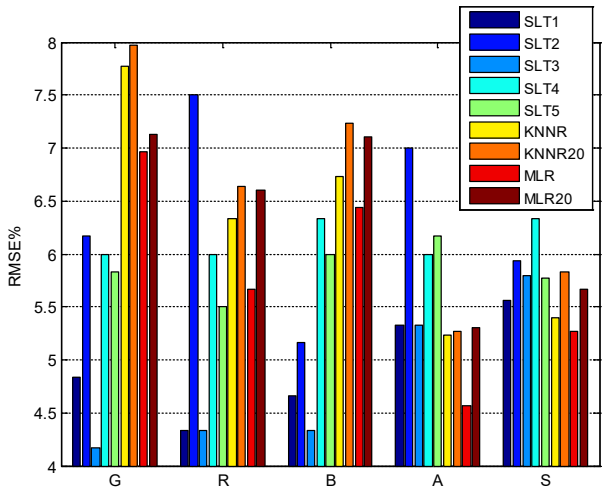


Figure 1: RMSE% for SLT 1-5, KNNR & MLR with feature selection and using all available 20 features (KNNR20 & MLR20).

A comparison of the GRBAS scoring produced by the five SLTs and the KNNR and MLR algorithms is presented in Figure 1. This graph summarises the results of experiments carried out by Jalalinajafabadi [14] with and without feature selection. Measurements obtained without feature selection are labelled KNNR20 and MLR20 since all available 20 features are taken into account. Comparing KNNR (with feature selection) and KNNR20, the feature selection has reduced the prediction error RMSE% by up to about 0.5%. Comparing MLR and MLR20, the reduction due to feature selection is generally greater, i.e. about 1% for Roughness and up to 0.7% for the other components (apart from Grade). With feature selection, the performances of the two machine learning techniques appear quite similar according to the RMSE measurements, though MLR is consistently better than KNNR. For Asthenia and Strain, both KNNR and MLR with feature selection deliver a lower RMSE than was obtained for each of the SLT raters with reference to the corresponding reference-

scores. For Grade, the KNNR and MLR values of RMSE (with feature selection) are both markedly higher than the corresponding values obtained for all the five SLT raters. The worst RMS difference for Grade is about 7.5%. The results for 'Breathiness' are close to those of the two worst performing SLT raters, and the MLR result for 'Roughness' lies between the two best and two worst performing SLT raters. As reported by Jalalinajafabadi [14] and further explained in [1], the RMSE taken over all GRBAS components was found to be marginally lower for KNNR and MLR (both with feature selection) than for each of the five individual SLT raters.

14. Conclusions

Recordings of normal and impaired voices were obtained from randomly selected patients and some other volunteers. These recordings were audio-perceptually assessed by five expert GRBAS raters to obtain a set of GRBAS scores for each recording. Statistical methods for measuring the inter-rater and intra-rater consistency of the scoring were investigated and it was concluded that the linearly weighted Cohen Kappa (*LwCK*) was suitable for this purpose. The measurements suggested that the GRBAS assessments were reasonably consistent. The scores and *LwCK* consistency measurements were then used to produce a set of 'reference scores' for training machine learning algorithms for mapping acoustic feature measurements to GRBAS scores, and thus performing automatic GRBAS scoring. With the reference scores, and acoustic feature measurements extracted from each of the 102 speech recordings by standard DSP techniques, KNNR and MLR were found to produce comparable automatic GRBAS scoring performances which compared favourably with the scoring by the five SLT raters. Feature selection was applied to determine the best subset of the twenty available acoustic features for each GRBAS dimension.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgment

The authors acknowledge the contributions of Ms Frances Ascott, the SLT raters and the participants. We also acknowledge considerable help and advice from Prof. Gavin Brown and Prof. Mikel Lujan in the Computer Science School of Manchester University, UK.

References

[1] Z. Xie, C. Gadepalli, F. Jalalinajafabadi, B.M.G. Cheetham, J.J. Homer, "Measurement of Rater Consistency and its Application in Voice Quality Assessments", 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, Shanghai, China, October, 2017.

[2] M. Hirano, "Clinical Examination of Voice", New York: Springer, 1981.

[3] M.S. De Bodt, F.L. Wuyts, P.H. Van de Heyning, C. Croux, "Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality", Journal of Voice, 1997, 11(1):74-80.

[4] C. Sellars, A.E. Stanton, A. McConnachie, C.P. Dunnet, L.M. Chapman, C.E. Bucknall, et al., "Reliability of Perceptions of Voice Quality: evidence from a problem asthma clinic population", J. Laryngol Otol., 2009, pp. 1-9.

[5] A.L. Webb, P.N. Carding, I.J. Deary, K. MacKenzie, N. Steen & J.A. Wilson, "The reliability of three perceptual evaluation scales for dysphonia", Eur Arch Otorhinolaryngoly; 261(8):429-34, 2004.

[6] J. Laver, S. Wirz, J. Mackenzie & S. Hiller, "A perceptual protocol for the analysis of vocal profiles", Edinburgh University Department of Linguistics Work in Progress; 14:139-55. 1981.

[7] D. K. Wilson, "Children's voice problems", Voice Problems of Children, 3rd ed., Williams and Wilkins, Philadelphia, PA., 1-15, 1987.

[8] G.B. Kempster, B.R. Gerratt, K.V. Abbott, J. Barkmeier-Kraemer & R.E. Hillman, "Consensus Auditory-Perceptual Evaluation of Voice: Development of a Standardized Clinical Protocol", American Journal of Speech-Language Pathology; 18(2):124-32, 2009.

[9] P. Carding, E. Carlson, R. Epstein, L. Mathieson & C. Shewell, "Formal perceptual evaluation of voice quality in the United Kingdom", Logopedics Phoniatrics Vocology, 25(3):133-8, 2000.

[10] S.N. Awan, and N. Roy, "Toward the development of an objective index of dysphonia severity: a four-factor acoustic model", Clinical linguistics & phonetics, 20(1):35–49, 2006.

[11] S.N. Awan, N. Roy, M.E. Jette, G.S. Meltzner and R.E. Hillman, "Quantifying dysphonia severity using a spectral/cepstral-based acoustic index: comparisons with auditory-perceptual judgements from the CAPE-V", Clinical Linguistics & Phonetics, 24(9):742–758, 2010.

[12] T. Bhuta, L. Patrick & J.D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements", Journal of Voice, Elsevier, 2004, Vol.18, Issue.3, pp. 299-304.

[13] F. Villa-Canas, J.R. Orozco-Arroyave, J.D. Arias-Londono et al., "Automatic assessment of voice signals according to the GRBAS scale using modulation spectra, MEL frequency cepstral coefficients and noise parameters", IEEE Symposium of Image, Signal Processing, and Artificial Vision (STSIVA), 2013, pp. 1-5.

[14] F. Jalalinajafabadi, "Computerised assessment of voice quality", PhD Thesis. 2016, University of Manchester, UK.

[15] J. Lee Rodgers & W.A. Nicewander, "Thirteen ways to look at the correlation coefficient", The American Statistician, 1988, vol. 42(1), pp. 59-66.

[16] J.F. Bland and D.G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement", The lancet, 1986, 327(8476), pp. 307-310.

[17] J. Krieman, B.R. Gerratt, G.B. Kempster, A. Erman & G.S. Berke, "Perceptual Evaluation of Voice Quality: Review, Tutorial and a Framework for Future Research", Journal of Speech and Hearing Research, Vol. 36, 21-40, 1993, pp 21-40.

[18] G.G. Koch, "Intraclass correlation coefficient", Encyclopedia of statistical sciences, 1982.

[19] J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychosocial Measurement, 1960, 20, pp. 37-46.

[20] J. Cohen, "Weighted Kappa: Nominal scale agreement provision for scaled disagreement or partial credit", Psychological bulletin, 1968, vol. 70(4), p. 213.

[21] J.L. Fleiss, "Measuring nominal scale agreement among many raters, Psychological bulletin", 1971, vol. 76 no 5, pp. 378-382.

[22] A.J. Viera and J.M. Garrett, "Understanding inter-observer agreement: the Kappa statistic", Fam Med. 2005, vol. 37(5), pp. 360-3.

[23] J.L. Fleiss, "Design and analysis of clinical experiments", Vol. 73. John Wiley & Sons, 2011.

[24] J.D. Evans, "Straightforward Statistics for the Behavioral Sciences", Brooks/Cole Publishing Company, 1996.

[25] E. Rödel & R.A. Fisher, "Statistical Methods for Research Workers", 14. Aufl., Oliver & Boyd, Edinburgh, London 1970. XIII, 362 S., 12 Abb., 74 Tab., 40 s. BiometrischeZeitschrift, 1971, vol. 13(6), pp. 429-30.

[26] J.L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intra class correlation coefficient as measures of reliability", Educational and Psychological Measurement, 1973, vol. 33, pp. 613-619.

[27] M.J. Warrens, "Inequalities between Multi-Rater Kappas", Adv Data Classif, 2010, vol. 4, pp. 271-286.

[28] R.J. Light, "Measures of response agreement for qualitative data: some generalisations and alternatives", Psychol Bull, 1971, vol. 76, pp. 365-377.

[29] L. Hubert, "Kappa Revisited", Psychol Bull 1977, vol. 84, pp. 289-297.

[30] A.J. Conger, "Integration and Generalisation of Kappas for Multiple Raters", Psychol Bull.,1980, vol. 88, pp. 322-328.

[31] Z. Xie, C. Gadepalli, & B.M.G. Cheetham, "A study of chance-corrected agreement coefficients for the measurement of multi-rater consistency", International Journal of Simulation: Systems, Science & Technology 19(2), 2018, pp. 10.1-10.9.

[32] P. Boersma & D. Weenink, "Praat: a system for doing phonetics by computer", Glot International (2001) 5:9/10, pp. 341-345.

[33] O. Amir, M. Wolf & N. Amir, "A clinical comparison between two acoustic analysis softwares: MDVP and Praat", Biomedical Signal Processing and Control, 2009, vol.4(3), pp. 202-205.

[34] S. Hadjitodorov & P. Mitev, "A computer system for acoustic analysis of pathological voices and laryngeal diseases screening", Medical engineering & physics, 2002, vol. 24(6), pp. 419-29.

[35] F. Jalalinajafabadi, C. Gadepalli, F. Ascott, J.J. Homer, M. Luján & B.M.G. Cheetham, "Perceptual Evaluation of Voice Quality and its correlation with

acoustic measurement", IEEE European Modeling Symposium (EMS2015), Manchester, 2013, pp. 283-286.

[36] F. Jalalinajafabadi, C. Gadepalli, M. Ghasempour, F. Ascott, J.J. Homer, M. Lujan & B.M.G. Cheetham, "Objective assessment of asthenia using energy and low-to-high spectral ratio", 2015 12th Int Joint Conf on e-Business and Telecommunications (ICETE), vol. 6, pp. 576-583, Colmar, France, 20-22 July, 2015.

[37] F. Jalalinajafabadi, C. Gadepalli, M. Ghasempour, M. Lujan, B.M.G. Cheetham & J.J. Homer, "Computerised objective measurement of strain in voiced speech", 2015 37th Annual Int Conf of the IEEE Engineering in Medicine and Biology (EMBC), pp.5589-5592, Milan, Italy, 25-29 Aug 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 339-346 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

MRI images Enhancement and Brain Tumor Segmentation

Aye Min[*,1], Zin Mar Kyu[2]

[1]*Digital Image Processing, University of Computer Studies (UCSM), 05013, Myanmar*

[2]*Software Department, University of Computer Studies (UCSM), 05013, Myanmar*

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 03 November, 2018
Online: 01 December, 2018

Keywords:
MRI
Fusion based results binding
Adaptive K-means clustering

ABSTRACT

Brain tumor is the abnormal growth of cancerous cells in Brain. The development of automated methods for segmenting brain tumors remains one of the most difficult tasks in medical data processing. Accurate segmentation can improve diagnosis, such as evaluating tumor volume. However, manual segmentation in magnetic resonance data is a laborious task. The main problem to detect brain tumors is less precise to determine the area of the tumor and determine the segmentation accuracy of the tumor. The system proposed the fusion based results binding for MRI image enhancement and combination of adaptive K-means clustering and morphological operation for tumor segmentation. BRATS multimodal images of brain tumor Segmentation Benchmark dataset was used in experiment testing.

1. Introduction

Primary diagnosis of brain tumors is extremely significant, because it can save lives. Accurate segmentation of brain tumors is also important, as it can help medical personnel in the planning of treatment and intervention. Manual segmentation of tumors requires a long period of time, even for a qualified specialist. Fully automated segmentation and quantitative analysis of tumors, therefore, are highly beneficial maintenance. However, it is also very difficult due to the large variety of anatomical structures and low contrast of current imaging techniques that create the distinction between normal and tumor regions. The main objective of our research is to create a trustworthy procedure detection of tumors of a multimodal MRI record based on a controlled machine study the methods using a data set containing MICCAI Brats images with ground truth, provided by human experts.

In this article we propose fusion-based results binding (result fusion) method for image enhancement and combination of adaptive k-means clustering and morphological operation for tumor segmentation and reliable detection system are proposed. In MRI image enhancement, the definition of results fusion method is fusion the filtered results of median filter and wiener filter. We emphasized on the results of median filter and wiener filter. Median filter made the original image to be more sharpening and wiener filter made the image to be more smoothness. Segmentation of brain tumors is extremely significant, because it can save lives. Accurate segmentation of brain tumors is also important.

By fusion of these results, we got more sharpening and more smoothness image in this research. This is one of contribution for our article. Second is adaptive K-mean clustering is used like as segmentation method in this article. And then, we proposed the usage of opening and closing in morphological operation in this research. Second contribution of our article is the combination of clustering method and modified morphological operation to segment the MRI images. The experimental results will be discussed in the next chapters. This article is implemented with 6 chapters. Chapter 1 is Introduction, chapter 2 is state of the art, chapter 3 is theory background, chapter 4 is material and method, chapter 5 is experiments and chapter 6 is conclusion of the paper. This article is extended version of our paper in PDCAT'17 conference with the title of "MRI images Enhancement and Brain Tumor Segmentation" [1]. In this article, more datasets were tested and presented about the kernel of filters, discussed about the proposed system details.

2. State of The Art

S. Jeevakala and B. Therese described the paper title with "Non Local Means Filter Based Rician Noise Removal of MR Images" in 2016. In this paper, the authors proposed a combination of NLM and stationary wavelet transform (SWT) with adaptive thresholding to remove Rician noise and preserve structural information of edges. The proposed noise elimination algorithm will be useful for the subtle analysis of tissue / organ images [2].

[*]Aye Min, UCSM, +959789811977 & ayemin@ucsm.edu.mm

M. N. Nobi and M. A. Yousuf proposed the paper title with "A New Method to Remove Noise in Magnetic Resonance and Ultrasound Images". The proposed method is compared with a smoothing, medium and midpoint filter using quantitative parameters such as PSNR, SNR, and RMSE. The smoothing filter shows better results, but it is painful because of the blurring effect. In the medium filtering technique, it is considered that each pixel calculates the average and all the pixels are replaced by the calculated average. Therefore, the affected pixels are taken into account to calculate the average, and the unaffected pixels are replaced by this calculated average [3]. B. Shinde and AR Dani have announced a "Noise Detection and Removal Filtering Techniques in Medical Images" in 2012. In this experiment, various medical images, such as MRI, cancer, X-ray, brain, etc. All these medical images, after detection of Gaussian noise, use median filtering techniques to remove noise. The results they have achieved are more useful and found useful for general practitioners to easily analyze the patient's symptoms [4].

A. Mihailova et al. (2016) proposed the paper "Comparative Analysis Various Filters for Noise Reduction in MRI Abdominal Images. Gaussian noise is random noise, and has a normal distribution of the probability density function (also known as a Gaussian distribution). Rician noise is not additive noise, but it depends on the data. The median filter performs better than the Gaussian filter. Wiener filter works best, but the most significant results they get from the seismic pulse and especially the wavelet of the homomorphic filter [5]. The next paper is "Propagated Image filtering" and it was presented by J.H.R. Chang et al (2015). In this document, authors proposed a propagation filter as a local filtering operator with the objective of smoothing images while maintaining the context information of the image. Authors also propose technologies when propagation filtering is related to the propagation of beliefs and a greater acceleration of the filtering process is required. In the experiments, the propagation filter was applied to various applications, such as image noise reduction, smoothing, melting, high dynamic range (HDR) compression. Finally, several applications Computer vision and graphics have verified the effectiveness of propagation filters that have proven to be superior to existing image filters in both quantitative and qualitative assessments [6].

The another paper of image fusion is "Image Fusion using NSCT Theory and Wavelet Transform for Medical Diagnosis", authors are P. J. Anju et.al. There are different methods for medical image fusion. NSCT based fusion is further enhanced for better quality by integrating with wavelet fusion. The experimental results are tested and compared with other fusion methods by using the Peak Signal to Noise Raito (PSNR) and Structured Similarity Index Measure (SSIM) [7]. Md. Sujan et. al. (2016) proposed " A Segmentation-based automated system for the detection of brain tumors". In this work, we proposed threshold processing and a morphological method for detecting a tumor. Authors compared the results of the proposed method with the color segmentation method. They recognized the tumor area by comparing the true positive rate of segmented results. All results are tested with 72 flair sequences of the BRATS dataset [8]. Jyothsna et. al. (2015) proposed "Adaptive K-means clustering for Medical Image Segmentation". A number of the group's researchers focused on improving the clustering process. The proposed method of promoting the adaptive method is that clusters grow without first selecting the elements that constitute the cluster. It was discovered that it is capable of segmenting the region of different distribution intensity smoothly.

The method was used to achieve a significant process of accelerated research. In this paper, an adaptive clustering algorithm for K-tools is presented and does not depend on seed selection to initialize cluster K. The algorithm is tested for various images and works smoothly, resulting in good data separation and research resulting in the data structure being accelerated remarkably. One can conclude that adaptive to means that it works better, and the speed of K means [9].

Bobotov´ et al. (2016) proposed the title with "Segmentation of Brain Tumors from Magnetic Resonance Images using Adaptive Thresholding and Graph Cut Algorithm". They got the result of comparison Graph cut result and result without Graph cut [10]. The following article is described by Edily et al. entitled with "Detection and localization of brain tumors in magnetic resonance". The present inventors propose an automatic frame for the detection and localization of brain tumors capable of detecting and locating brain tumors in magnetic resonance images. The framework for detection and location of brain tumors proposed involves five steps: image acquisition, preprocessing, detection of edges, grouping of modified histograms and morphological manipulation. After morphological manipulation, the tumor appears as a pure white color on a pure black background. This system reached an error rate of 8%. Preliminary results demonstrate how a simple automatic learning classifier with a simple set of image-based features provides high classification accuracy. The preliminary results also demonstrate the effectiveness and efficiency of our 5-step approach to brain tumor detection and detection and extend this framework to detect other types of tumors in other types of medical images and motivate them to be localized [11].

Cabria et al. (2015) proposed "Automated Localization of Brain Tumors in MRI Using Potential-K-means Clustering Algorithm". In this paper, they viewed the intensity of a pixel as equal to its "workload" and employed an unsupervised learning algorithm called potential-K-means that generates a balanced distribution of the pixels into clusters of approximately equal total intensity. A set of 22 images of the FLAIR MRI (axial plane) modality from the BRATS dataset was used [12].

S. Priyanka, Dr. AS Naven kumar proposed the name of the noise elimination document "Noise Removal in Remote Sensing Image Using Kalman Filter Algorithm " in 2016. Remote sensing, in general, using sensors installed on airplanes and space platforms, the authors discussed Gaussian noise and speckle noise (salt and pepper). In this proposed study, authors reduced image noise using the Kalman filter and the Wiener filter. The Kalman filter is suitable for reducing noise while maintaining the basic structure of the image compared to other filters. The Kalman filter shows more filters with improved noise efficiency [13]. The next document is "An interactive graph cut method for brain tumor segmentation ", which is described by N. Birkbeck et al. We have developed a interactive semiautomatic brain tumor segmentation system that incorporates interactive 2D tools and automated 3D Propose Control. The method provided is based on the energy that incorporates the available MRI modalities and the regional

statistics calculated in the normal normalization period. The improvement of the new parameters includes the adjustment of the continuous balance of the adhesion parameters of the operator control and the user interaction in 2D line using Rasso and the brush tool (not including the point and plot click used in the segmentation previous interactive) There is. This improves segmentation control by drastically changing the statistics of the region and limiting the segmentation. Experiments have shown that the proposed tool accelerates segmentation compared to traditional manual segmentation and reduces reproducibility between users and users [14].

S. Priyanka, Dr. AS Naven kumar proposed the name of the noise elimination document "Noise Removal in Remote Sensing Image Using Kalman Filter Algorithm" in 2016. Remote sensing, in general, using sensors installed on airplanes and space platforms, the authors discussed Gaussian noise and speckle noise (salt and pepper). In this proposed study, authors reduced image noise using the Kalman filter and the Wiener filter. The Kalman filter is suitable for reducing noise while maintaining the basic structure of the image compared to other filters. The Kalman filter shows more filters with improved noise efficiency [13]. The next document is "An interactive graph cut method for brain tumor segmentation ", which is described by N. Birkbeck et al. We have developed an interactive semiautomatic brain tumor segmentation system that incorporates interactive 2D tools and automated 3D Propose Control. The method provided is based on the energy that incorporates the available MRI modalities and the regional statistics calculated in the normal normalization period. The improvement of the new parameters includes the adjustment of the continuous balance of the adhesion parameters of the operator control and the user interaction in 2D line using Rasso and the brush tool (not including the point and plot click used in the segmentation previous interactive) There is. This improves segmentation control by drastically changing the statistics of the region and limiting the segmentation. Experiments have shown that the proposed tool accelerates segmentation compared to traditional manual segmentation and reduces reproducibility between users and users [14].

3. Theory Background

3.1. Median Filter

The median filter is a non-linear method used to eliminate noise from the MRI brain images. And it is especially effective for eradicate salt and pepper noise. The median filter works by scrolling the pixel of the image with a pixel, replacing each value with the median value of the neighboring pixels. Pixels are calculated from the first sorting of all pixel values of adjacent patterns in the order, and then replace the pixel when viewed with a half pixel value. The median filter is capable of eliminating noise without degrading the sharpness of the image [5].

$$y[m,n]=\text{median}\{x[i,j],(i,j)\varepsilon\omega\} \quad (1)$$

Where ω is a neighborhood defined by the user, centered around location [m,n] in the image. An example of median filter of 3*3 kernel or window size is shown below. We take the original values and order the values to 0, 2, 3, 3, 4, 6, 10, 15 and 97. We find the medium value and fill this value to the center point. So, centered value 97 is replaced by the medium of all nine values 4.

Unfiltered value

*	*	*
*	4	*
*	*	*

Filtered value

6	2	0
3	97	4
19	3	10

3.2. Wiener Filter

Anti-aliasing is the Wiener (non-linear) filter. This filter simultaneously eliminates noise and blurry integrals. There are two parts of the work: the inverse filter and the noise leveling. The Wiener filters are a class of optimal linear filters, with noisy data because the inputs are the calculation of the difference between the output sequences required of the actual output. Performance supervision can be considered the error of least squares. There is also a Wiener2 filter is an adaptive 2-D noise removal filter. This function works as a filter application. Wiener is a type of image for a linear adaptive filter that adapts to the local variance of the image. Wiener2 has done little to smooth out the great variance. At the small wiener2, more sanding is lit. Therefore, it is often better than linear filtration. In comparison, an adaptive filter is more intuitive than a comparable linear filter, parts of the profiles and high-frequency images. There is no design work, the wiener2 function processes all preliminary calculations, preliminary calculations and filter equipment and the implementation of the general input filter. The Wiener2 filter is more suitable for repairing Gaussian noise.

3.3. Image Fusion

Image fusion is a method of integrating all images of applicable information and balanced similar sources or multiple sources in one merged image without any degradation. The main goal of merging medical imaging is the reliable integration of the observation from different input images into one image, not including any degradation and loss of visual information. There were three main ways of fusion of images - pixel level, feature level and decision level. The pixel level is a low fusion of images, the address of the pixels obtained at the output of the image sensor. Fusion of images at pixel level refers to a mixture of information and synergistic data collected from various image sources to provide an improved type of view. When the merging of images is combined at the pixel level directly into the information layer, the amount of information is greater. Almost all the image algorithms of the merged ones are designed to fall at the pixel level [6].

3.4. Adaptive K-Mean Clustering

Clustering is a major problem in a wide range of fields, such as pattern recognition and artificial vision. The general grouping method is based on the average K. However, it has four main drawbacks. First, it is a late and incorrect time scale. Second, it is often not desirable to wait for the user to identify the number of clusters. Third, it can exacerbate excellent areas. Finally, its performance depends to a large extent on the initial center of the cluster. To overcome the above drawbacks, the grouping algorithm 4 in this document means that K (AKM) is effectively adapted. The AKM to estimate the correct number of clusters and

obtain the initial segmentation of the histogram in the center of the linear norm with the linear norm is composed of a set of data and then a local heuristic improvement to group the K means to avoid the optimal values. Execute the algorithm In addition; the kd tree is used to store data sets to accelerate. AKM has been tested on synthetic and real image data sets.

3.5. Morphological Operation

In morphological operations, binary images and structural elements are often used as input and in combination with use of switches (intersections, conjunctions, inclusions, complements). The processed image object is based on the characteristics of the shape encoded by the input structure element. Mathematical details are described in mathematical morphology. Each pixel of the image, as well as each pixel of the whole image, is compared to a set of elementary pixels. If the group of two elements corresponds to a condition defined by a group operator (for example, when a plurality of pixels of the structural element is a subset of the pixels of the basic image), the pixels below the origin of the structural element are determined Values (for binary images, prerequisites 0 or 1).

Opening. Opening eliminates small objects in the foreground (usually taken as a bright pixel) of images by placing in the background, while closing is eliminating small holes in the foreground and changing the background of small islets in the foreground. These methods can also be used to find specific paths in the image.

$$A^{o}B = (A \ominus B) \oplus B \quad (2)$$

Where $\ominus$ and $\oplus$ indicate erosion and dilation, respectively.

Closing. When processing images, closing together with the opening, the basic workhorse of morphological is to eliminate the noise. Opening is eliminating small items, while closing eliminates small holes.

$$A^{o}B = (A \oplus B) \ominus B \quad (3)$$

Where $\ominus$ and $\oplus$ indicate erosion and dilation, respectively.

4. Materials and Methods

4.1. Proposed Method

MRI images are posh by noise such as rice (Rician), Gaussian, salt and pepper. In order to confiscate noise, many noise filtering methods have been proposed. In this system, we propose a fusion method of results for improving MRI image. The way to fuse the results is built by fusion the results of the median filter with the results of the Wiener filter. Direct fusion method is used when merging results. The performance of the medium filter and the winning filter depends on the size of the core or the size of the window. We evaluate and select the best kernel in our study. We examined the choice of kernel size by testing three sets of MRI image data. They are the data set BRATS, DICOM, TCIA. The values of the kernel (3, 5, 7 and 9) or the size of the window are tested and analyzed. Figure (1) shows a part of the sample image of the data set tested for kernel analysis. In MRI images, noise elimination and kernel value [3 * 3] are very effective in processing time and reconstructed image quality.

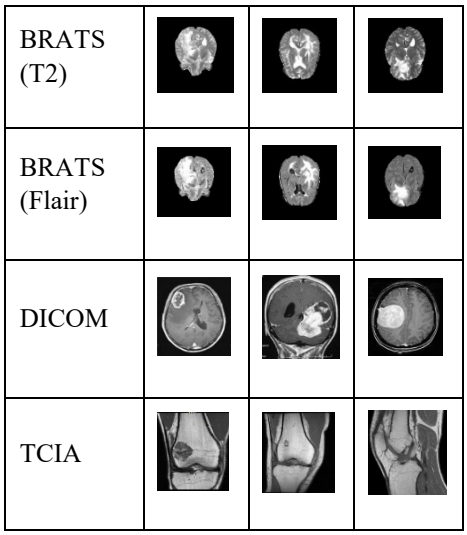


Figure 1: Sample images of datasets

Table 1: Average result of kernel analysis on T2

Kernels	Median Filter		Wiener Filter	
	RMSE	PSNR	RMSE	PSNR
3*3	5.506	33.499	2.677	39.646
5*5	8.946	29.221	3.91	36.35
7*7	11.368	27.107	4.966	34.265
9*9	13.175	25.810	5.905	32.757

Table 2: Average result of kernel analysis on flair

Kernels	Median Filter		Wiener Filter	
	RMSE	PSNR	RMSE	PSNR
3*3	5.567	33.645	2.782	39.557
5*5	8.971	29.483	3.918	36.511
7*7	11.044	27.608	4.648	34.974
9*9	12.623	26.406	5.294	33.822

Table (1), (2), (3) and (4) present the average values of all kernels on T2, Flair, DICOM and TCIA datasets respectively. Allowing to the results of the analysis, the kernel [3 * 3] is ideal for the elimination of noise and the emphasis processing of the

MRI images. Therefore, in this article, we select the median filter and the kernel value [3 * 3] of the Wiener filter.

Table 3: Average result of kernel analysis on DICOM

Kernels	Median Filter		Wiener Filter	
	RMSE	PSNR	RMSE	PSNR
3*3	4.516	35.306	3.888	36.607
5*5	9.965	28.352	6.996	31.428
7*7	15.301	24.580	9.410	28.827
9*9	20.171	22.150	11.324	27.213

Table 4: Average result of kernel analysis on TCIA

Kernels	Median Filter		Wiener Filter	
	RMSE	PSNR	RMSE	PSNR
3*3	8.096	31.086	5.269	34.248
5*5	5.077	34.052	4.637	34.841
7*7	6.672	31.680	5.771	32.940
9*9	7.898	30.214	6.483	31.929

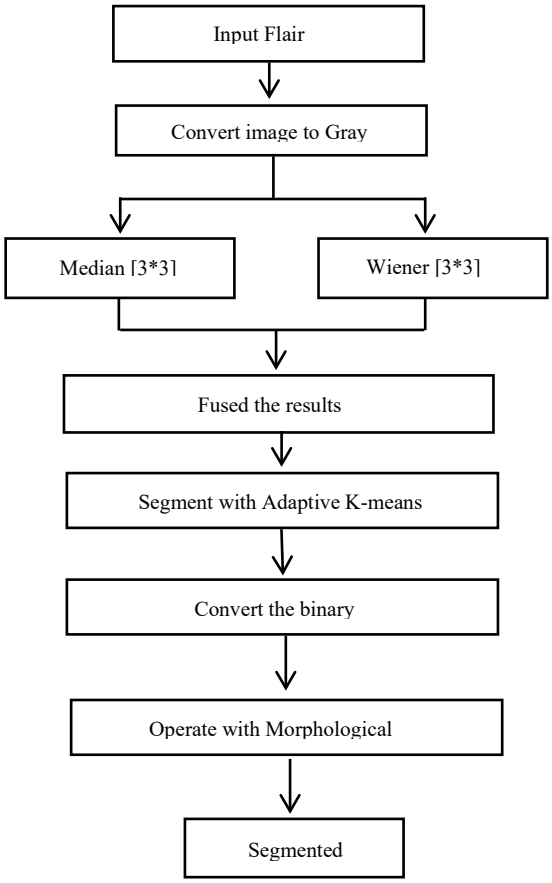


Figure 2: Overview of Proposed System.

Second proposed method is the combination of Adaptive K-means clustering and Morphological operation for MRI images segmentation. The system receives the RGB image and converts the RGB image into a grayscale image. Then, the grayscale image is filtered simultaneously by the medium filter and the Wiener filter. Medium and winning filters use the kernel value [3 * 3] or the size of the window to reduce noise. [3 * 3] The value of the kernel is more powerful and adequate to eliminate noise from MRI images [15]. Therefore, this kernel is used in this research document. Both filtered results are combined with image fusion. The merged image is segmented using adaptive k-means clustering, after which the segmented image is transformed into a binary image with a threshold of 0.7. The morphological operation re-segments the binary image. In this way, the operation closing and opening are used in order. Closing is [1; 1; 1; 1; 1; 1; 1] value of the kernel and the value of the kernel value Opening is [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]. After applying the morphological manipulation, the system generates images of the tumor segment.

5. Experiments

5.1. Image Enhancement Results

The proposed system is constructed by fusion the result of the Medan filter and the result of the Wiener filter. The results of the proposed method are compared with the Medium filter and the Wiener filter. The results of these methods are evaluated using the values of MSE (mean square error) and PSNR (peak signal noise ratio). The table (5) shows the average results of 40 images of Flair and T2. In figure (3), the design of results fusion method is described.

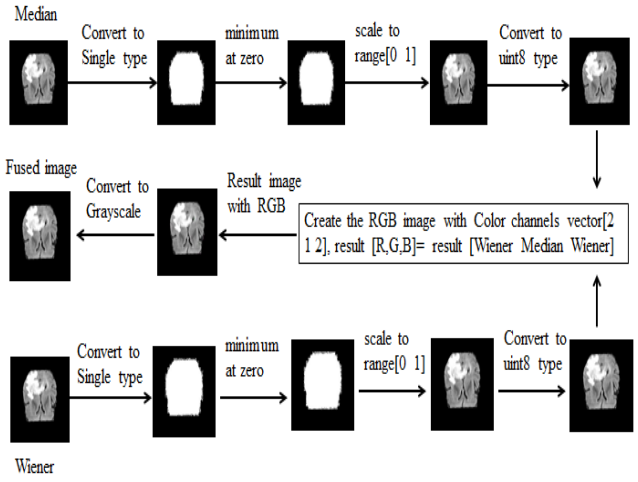


Figure 3: Design of the fusion-base results binding method

Table 5: Average result of MSE and PSNR on 40 flair and 40 T2

Method	**MSE**		**PSNR**	
	T2	***Flair***	***T2***	***Flair***
Median filter	28.6604	35.5557	34.1986	33.3925
Wiener filter	31.6925	32.9356	33.4274	33.2784
Proposed Method	21.9649	25.8498	35.1073	34.5308

Table 6: Average result of 72 flair images

Methods	TPR (%)	TNR (%)	PVP (%)	A (%)
AKM	58.47	99.34	87.21	96.16
AKMM (proposed)	85.41	98.90	78.30	98.30
Otsu	55.998	99.39	88.55	95.08
Region growing (RG)	73.64	99.28	86.50	97.42
Particle Swarm Optimization (PSO)	56.10	99.39	88.58	95.13
Interactive Graph Cut (IGC)	42.00	92.64	85.49	86.20

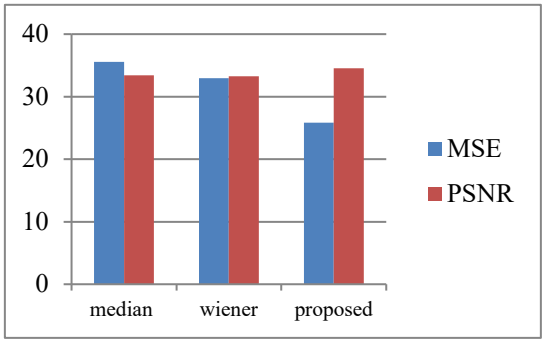


Figure 4: Overview of the Results of MSE and PSNR on T2

5.2. Image Segmentation Results

In this system, the combination of adaptive K-means clustering and morphological approach (AKMM) is proposed for tumor segmentation. Tumor segmentation was tested by two ways to detect advantages and disadvantages of proposed algorithm. First, it was tested with the algorithm which consists only of proposed results fusion method and adaptive K-means clustering (AKM) algorithm. Second, it was tested also with the morphological operation. There are two evaluation methods to use testing accuracy of tumor segmentation. First method includes True positive rate, True negative rate, Predictive value positive and Accuracy. Second method includes Jaccard Similarity index. Resulting tumor segmentation was divided into true positive (TP), true negative (TN), false positive (FP) and false negative (FN) regions. TP represents pixels where tumor was detected and it should be tumor. TN means that tumor was not detected and should not be. FP is when tumor was detected and should not be. Finally if tumor was not detected, but should be, it is FN. Statistical methods were used to evaluate results:

True positive rate – sensitivity (TPR):

$$TPR=TP/TP+FN \quad (4)$$

True negative rate- specific (TNR):

$$TNR=TN/TP+FN \quad (5)$$

Predictive value positive – precision (PVP):

$$PVP=TP/TP+FP \quad (6)$$

Accuracy (A):

$$A=TP+TN/TP+FP+TN+FN \quad (7)$$

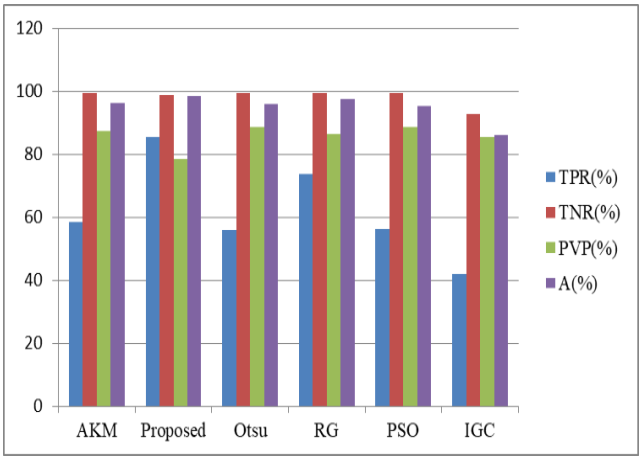


Figure 5: The Results of Comparement Method.

Table 7: Average result of jaccard similarity index in 72 flair images

Methods	Jaccard	Rfn	Rfp
AKM	0.527(min= 0.174, max=0.7996)	0.9953	1.44
AKMM (proposed)	0.6851(min= 0.3672, max=0.8427)	0.9958	1.011
Otsu	0.51(min= 0.0699, max=0.834)	0.9953	2.12
Region growing (RG)	0.647(min= 0.1124, max=0.87)	0.9954	1.594
Particle Swarm Optimization (PSO)	0.513(min= 0.0672, max=0.834)	0.9953	2.13
Interactive Graph Cut (IGC)	0.393(min= 0, max=0.840)	0.9954	2.437

Table 8: Run time duration of comparement methods

Methods	TPR (%)	TNR (%)	PVP (%)	A (%)
AKMM	78.66	99.03	79.81	97.9
PSOM	66.16	99.29	85.99	96.47
OTSUM	66.44	99.29	85.99	96.47

Jaccard similarity coefficient is a statistic used for comparing the similarity and diversity of sample sets. rfn is ratio of false negative and rfp is ration of false positive.

$$\text{Jaccard}(A,B) = |A \cap B| / |A \cup B| \quad (8)$$

$$\text{rfn} = B - |A \cap B| / B \quad (9)$$

$$\text{rfp} = A - |A \cap B| / B \quad (10)$$

Table 9: Performance and comparison analysis of proposed method in BRATs dataset

Performance and comparison analysis of proposed method BRATS brain tumor dataset			
Algorithm	MRI modalities	Approach	True positive rate
Md. Sujan, Nashid Alam(2016)	72-Flair	Thresholding and morphological processing	84.72%
Proposed	72-Flair	Adaptive k-mean clustering and morphological processing	85.41%

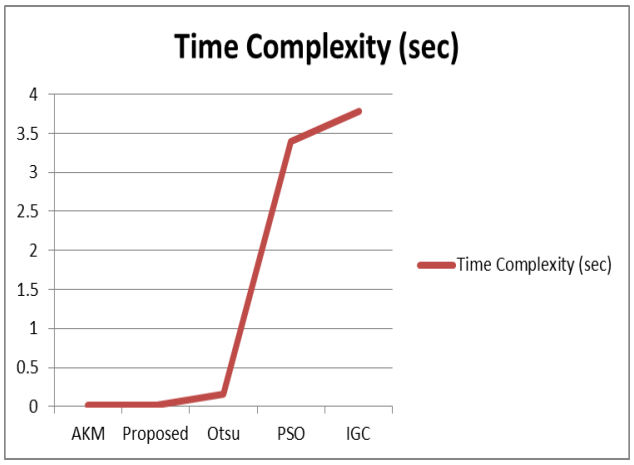


Figure 6: Time Complexity of Comparement Method

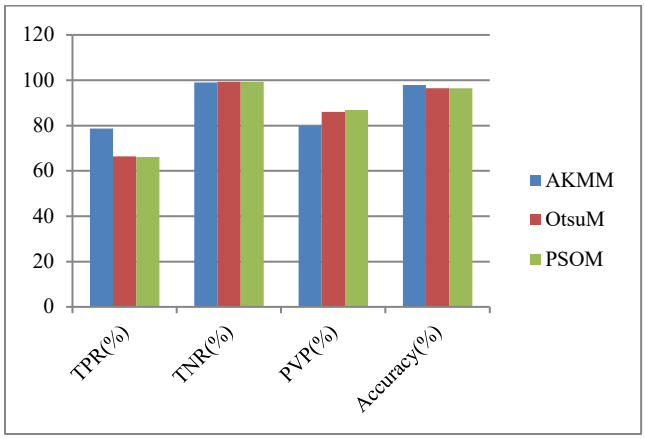


Figure 7: The Averages Results of 110 Flair Images

Table 10: Average results of 110 flair images

Method	Run-time (sec)
AKM	0.0106524
AKMM (proposed)	0.0192652
Otsu	0.15238
PSO	3.39533
IGC	3.77659

Experimental output of proposed Adaptive k-mean clustering and Morphological (AKMM) method				
Input image	Filtered image	Adaptive K-mean clustering	Morphological processing (output)	Ground Truth image

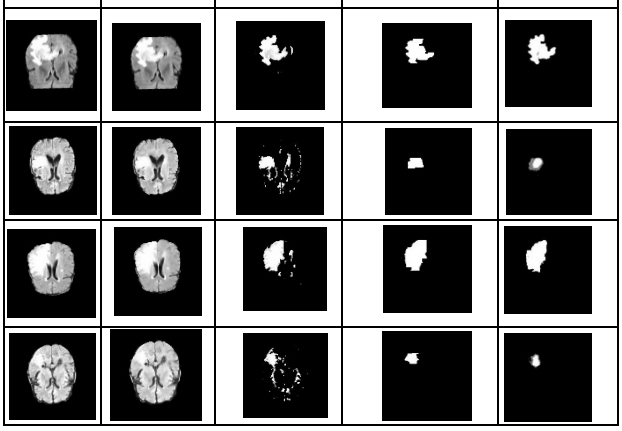


Figure 8: Experimental output of proposed Adaptive k-mean clustering and Morphological (AKMM) method

Table 10 describe about the average experimental results of 110 flair images. In this table, Our proposed method (AKMM) is compared with combination of Otsu and morphological operation and the combination of particle swarm optimization (PSO) and morphological operation. According to the results, our proposed method is more exceeding in Accuracy.

5.3. Research Discussion

In this article, we proposed tow contributions. First is fusion – based results binding to get better result of MRI images enhancement. In this method, we fused the results of Median filter and Wiener filter. At that time, we used more Wiener effect. Thus, our enhancement method image is more smoothness than original and we also save image sharpening from Median filter properties. Among the segmentation methods, almost the segmentation based segment methods are more suitable with median filter and almost the clustering base segment methods are more suitable with wiener filter. Our filtering and enhancement method can be used in both of segmentation. According to experimental results, our enhancement method got better results than Median filter and Wiener filter in MSE and PSNR values. So, we proposed first contribution is better results. And then, we proposed the second contribution.

Second contribution is combination of Adaptive K-means clustering and morphological operation (AKMM). We used proposed filter method in the preprocessing stage. First paper described the results based on 72 flair images. In this article, 110 flair images are tested and described. We also used the Jaccard similarity index in experimental results evaluations. Our proposed method (AKMM) got better results than other comparative base-line methods in accuracy and run time duration. All of the base-line methods are downloaded from matlab file exchange site and re-implemented by our self.

6. Conclusion

In this work, 40 flair and T2 sequences are tested for enhancement of MRI images and 110 flair sequences are tested for tumor segmentation. The fusion based results binding method was proposed for MRI images enhancement and combination of morphological operation and adaptive K-means (AKMM) for tumor segmentation. According to our experimental results, the proposed improvement is superior to Median filter and Wiener filter. Then we test and compare base-line methods such as Otsu's threshold, region growth, particle swarm optimization, and interactive graph cut segmentation with the proposed method (AKMM). The proposed method and all base-line methods are tested on 72 flair images and 110 flair images. The proposed method (AKMM) gained higher accuracy than the basic method compared, and it gets less complex over time.

References

[1] Aye Min and Zin Mar Kyu, "MRI images Enhancement and Tumor Segmentation for Brain", 18th International Conference on Parallel and Distributed Computing, Applications and Technologies, 0-7695-6330-9/17/31.00©2017IEEE DOI 10.1109/PDCAT.2017.00051

[2] S. Jeevakala, B. Therese, "Non Local Means Filter Based Rician Noise Removal of MR Images", International Journal of Pure and Applied Mathematics, Volume 109 No. 5 2016.

[3] M. N. Nobi and M. A. Yousuf, "A New Method to Remove Noise in Magnetic Resonance and Ultrasound Images", Journal of Scientific Research (JSR) Publication, 2011.

[4] B. Shinde and A.R. Dani, "Noise Detection and Removal Filtering Techniques in Medical Images", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 4, pp.311-316, July-August 2012.

[5] A. Mihailova, V. Georgieva, "Comparative Analysis Various Filters for Noise Reduction in MRI Abdominal Images", International Journal "Information Technologies & Knowledge" Volume 10, Number 1, © 2016.

[6] Rick Chang, J. H., & Frank Wang, Y. C., "Propagated image filtering" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 10-18) 2015.

[7] P. J. Anju, Dr.D.Loganathan," Image Fusion using NSCT Theory and Wavelet Transform for Medical Diagnosis", International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (3) , 2016.

[8] Md. Sujan, S.A. Noman, N. Alam, M. J. Islam,"A Segmentation based Automated System for Brain Tumor Detection", International Journal of Computer Applications (0975 – 8887) Volume 153 – No 10, November 2016.

[9] C. Jyothsna , Dr.G.R.Udupi, "Adaptive K-means Clustering for Medical Image Segmentation",International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Special Issue 31(September, 2015), PP. 15-21

[10] Z. Bobotov´, and W.S. Bene, "Segmentation of Brain Tumors from Magnetic Resonance Images using Adaptive Thresholding and Graph Cut Algorithm", The 20th Central European Seminar on Computer Graphics, Slovakia, Proceedings of CESCG 2016.

[11] Azhari, E. E. M., Hatta, M. M., Htike, Z. Z., & Win, S. L. "Brain tumor detection and localization in magnetic resonance imaging" International Journal of Information Technology Convergence and services (IJITCS), 4(1), 2231-1939 2014.

[12] I. Cabria, I. Gondra,"Automated Localization of Brain Tumors in MRI Using Potential-K-means Clustering Algorithm", IEEE, 12th Conference on Computer and Robot Vision 2015.

[13] S. Priyanka, Dr.A.S.N. kumar, " Noise Removal in Remote Sensing Image Using Kalman Filter Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016.

[14] Birkbeck, N., Cobzas, D., Jagersand, M., Murtha, A., & Kesztyues, T. "An interactive graph cut method for brain tumor segmentation". In Applications of Computer Vision (WACV), 2009 Workshop on (pp. 1-7). IEEE.

[15] Aye Min and Zin Mar Kyu, "Kernels analysis in MRI images Noise Removal Methods", 16th International Conference on Computer Application,ICCA(2018).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 347-357 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Semi-Autonomous Robot Control System with an improved 3D Vision Scheme for Search and Rescue Missions. A joint research collaboration between South Africa and Argentina

Jorge Alejandro Kamlofsky[*,1], Nicol Naidoo[2], Glen Bright[2], Maria Lorena Bergamini[1], Jose Zelasco[3], Francisco Ansaldo[3], Riaan Stopforth[2]

[1]*Universidad Abierta Interamericana (UAI), CAETI, C1270AAH, Buenos Aires - Argentina*

[2]*University of KwaZulu-Natal (UKZN), Mechatronics and Robotics Research Group (MR²G), 4000, Durban – South Africa*

[3]*Universidad de Buenos Aires (UBA), Facultad de Ingeniería, C1063ACV, Buenos Aires - Argentina*

ARTICLE INFO

Article history:
Received: 03 July, 2018
Accepted: 26 October, 2018
Online: 01 December, 2018

Keywords:
Search and rescue robots
3D vision
Robotic vision
Robotic navigation

ABSTRACT

Rescue operations require technology to assist the rescue process. The robotic technology in these missions is becoming very important. The important aspects investigated in this study are the integration of a mechatronic system that will allow for a robotic platform with a vision system.

The research collaboration between Argentina and South Africa is discussed, with the correlating research areas that each country investigated. The study permitted the development and advancement of a search and rescue system for different robots (wayfarer and drones) with different vision capabilities. A novel and innovative vision approach is presented.

1. Introduction

This paper is an extension of work originally presented in the 24th edition of the Mechatronics and Machine Vision in Practice (*M2VIP*) [1].

Search and rescue (SAR) is a general field referring to an emergency response to locate, give aid, medical care and to rescue people. Urban SAR (USAR) is considered a multiple-hazard discipline [2]. Urban environments are often more susceptible to human induced disasters: major industrial and/or transport accidents and/or fires, terrorism, wars, etc. or can have enormous effects in case of extreme natural phenomena like earthquakes, hurricanes, etc.

SAR operations in dangerous, disaster or catastrophe areas can be greatly improved with the assistance of tele-operated robots and/or semi-autonomous robots. Current implementations of mobile robots for SAR operations require human operators to control and guide the robot remotely. Even if human operation can be effective, operators may become stressed and tired rapidly [3]. In addition, current autonomous robots are incapable to work properly by moving in complex and unpredictable scenarios. This challenge could be addressed with a correct balance between the level of autonomy of the robot and the level of human control over the robot [4].

The first phase in a rescue mission is to identify the target area. The goal is to obtain a precise evaluation of the number and location of victims, detect dangerous situations for rescue personnel or survivors such as gas leaks, live wires, unsafe structures, etc. Semi-autonomous robots carrying correct sensors can obtain data from the field without risks of SAR personnel, and (if connected) can report on line important information. In the other hand, because time is critical during search and rescue operations in catastrophe areas, personnel do not have to waste time waiting for initialization and preparation for robots in the field. The robots must be deployed as soon as possible; if not, the acceptance of the technology by rescue personnel could prove difficult [3].

The mining activity in South Africa has a long list of victims of accidents. This is a case among others, where semi-autonomous mobile robots can access into unstable and toxic areas, in mine shafts, in order to identify dangerous areas, and to save accident victims and rescuers [3]. Argentina had also events in which the use of semi-autonomous robots could have been useful in SAR operations: terrorist attacks, mudslides, and the recent

[*]Jorge Alejandro Kamlofsky, Email: jorge.kamlofsky@gmail.com

disappearance in the ocean of the military submarine ARA San Juan.

The development and deployment of technology associated with SAR robots will also have a direct benefit to people who live in poor neighbourhoods of South Africa, and Argentina, and in many other circumstances. Fires and mudslides can have a devastating effect in low cost housing developments and the SAR of victims can be improved with technologically advanced semi-autonomous SAR robots. In addition, SAR operations in other disaster cases like, floods [5], mudslides [6], earthquakes, avalanches, fallen buildings, nuclear fusions, explosions, terrorist attacks, forest and/or industrial fires, recovery and deactivation of bombs, etc. can be greatly enhanced with the aid of semi-autonomous robots [7].

SAR personnel use semi-autonomous robots to improve SAR operations. These robots help to recover wounded or trapped people in disaster areas that are dangerous or can pose a threat to human rescuers. SAR operations using semi-autonomous robots can be carried out faster without risk for the SAR personnel. An advantage of using semi-autonomous robots in SAR operations is that they are not susceptible to toxic agents as gases, radiation, acidic or alkaline spills, etc. where toxicity could be dispersed. So, they can access to areas where rescuers cannot reach, or they can do it by risking their lives. If an accident occurs during the operation, there is no injury or loss of human life.

The collapse of the buildings of the Twin Towers in New York (USA) [3], recent catastrophes such as the Fukushima nuclear disaster in 2011 or Nepal earthquake in 2015 showed that SAR or USAR robots can be used to efficiently support rescue teams in finding persons in danger or gathering information more effectively. In Fukushima emergency responders deployed a SAR robot, to check on conditions in the surroundings and allowed workers a safe distance from hazardous radiation [8].

It is proven that semi-autonomous robots are a very useful tool for SAR operations showing the need to investigate how to improve the capacity of SAR robots around the world.

The objective of the joint collaboration of scientific and technological research between South Africa and Argentina is to research, design, develop, assemble and test different systems to improve the functionality of semi-autonomous robots for SAR operations and to improve the quality of life of inhabitants by carrying out rescue enhanced SAR missions in South Africa and Argentina.

The main objective of this paper is to discuss and to analyze the design and development of a semi-autonomous robotic platform that can be implemented in SAR missions in order to improve the results of the operations, as part of the joint research collaboration between South Africa and Argentina. An underlying objective is to remark the importance of the need of research and development of semi-autonomous robotics in order to assist to SAR operations. As an interesting result, in this paper, a simple and promising approach to 3D vision is presented.

The structure of this paper is as follows: Section 2 discusses the methodology of the research collaboration between South Africa and Argentina; Section 3 identifies the basic components (building blocks) of navigation and local control in robots and discusses the need for a middle-ware service. Section 4 presents the different vision algorithms, techniques and hardware used: the vision system, called here: "The Vision-Ware". Section 5 visualizes the design of a robot control architecture for SAR applications that will be integrated with the middle-ware. Section 6 discusses conclusions and future works.

2. Methodology

The methodology consists of research, design, simulation, development, assembly and testing of different subsystems for semi-autonomous mobile robots in SAR operations. The objective of the research collaboration is to research and develop innovative solutions related to vehicle chassis design, propulsion, navigation, guidance and orientation and 3D vision of the robot.

The methodology will allow the development of new technologies that can optimize the performance of robots in hostile environments so that vehicles can perform the necessary SAR operations reducing risks for human lives.

The research methodology will also focus on the study of new technologies that can be integrated on board of the semi-autonomous robot in order to deliver the necessary first aid equipment and to recover data from field. The methodology includes the development of simulations that will show the viability of subsystems developed and the performance of the robots in different tasks.

The University of KwaZulu-Natal has integrated a drone (Phantom 3) and a prototype platform (Segway RMP400) for vehicle design and propulsion technologies for semi-autonomous mobile robots. Further development and control of this system will be researched and investigated to allow semi-autonomous and autonomous control for SAR purposes.

Argentinean researchers will continue with the investigations carried out in mono and stereo vision, and the integration of cameras, which will help with the location of victims in a disaster environment and navigation and control of the semi-autonomous robotic platform.

3. Background

3.1. Brief and Trends of SAR Robotics

There is a vast literature related development of SAR Robots. Some papers that can show trends are mentioned bellow.

Actually, because of the most SAR robots are tele-operated, and autonomy robotics do not fit conditions for work and transit in complexity of SAR scenarios, a strong trend is the research and development of semi-autonomous robots, trying to obtain a good balance between the level of autonomy of the robot and human control over the robot [4].

Area mapping and localization for SAR missions research topic since many years, made good progress in terms of performance. A low bandwidth radar-based scanning-technology for mapping is presented in [9]. The radar technology is capable of showing a substantial mapping quality [8].

Swarm robotics consists of a large number of simple and tinny robots working together that perform tasks of greater complexity. It is strongly focused in coordination, cooperation and

collaboration. In particular for swarm robotics applied to SAR operations, optimization algorithms for hazardous environments and laborious tasks were proposed and development [10].

In [11], a case study of IOT application for SAR operations based on cloud is presented. A large-scale deployment of IOT devices in catastrophe scenarios controlled by a cloud based application implementing a "Robot-as-a-Service" scheme (RaaS) giving to SAR operations the advantages of cloud schemes applied to the use of robotic resources: flexibility, cost-efficient, scalability, and virtualization, between others.

3.2. Local Navigation and Control

A very important goal of long term research is the development of a robust navigation area mapping system for mobile robots for unknown and changeable scenarios [12].

The following building blocks of navigation are identified [12, 13]:

- Perception: the robot must obtain field data (raw-data) from their sensors and interpret it and convert it into useful information. Low-level sensing processes extract basic features such as highness or line segments, while high level sensing processes use symbolic models or geometric templates to constrain the sensor interpretation.
- Localization: The robot must identify its position in the field. To achieve this, the robots must have a location support system such as a GPS. If it is not possible, they must recognize the environment in order to know their position on the map. Even more: in several disaster areas, scenarios can vary from saved information. Autonomous vehicles must rely on information obtained from sensor data. They have to be integrated into a unified and consistent mapping model. The positional drift of the sensors due to the robot motion has to be taken into account in the mapping and navigation procedures.
- Cognition: the robot has to decide how to achieve its objectives in the target zone by planning a convenient path (the shortest one in a specific sense).
- Motion Control: In order to move over a planned path, the robot must control and drive its engines connected to its traction system.

In addition to these building blocks, robots must avoid any obstacle, static or dynamic, that may be present on the way to the goal. And if this is not possible, they should be able to plan a new path on the fly. Figure 1 shows a diagram containing the necessary components for the design of semi-autonomous robot system. Regarding to remotely controlled mobile robots, tasks like location, sensing and perception and a part of the cognition module must be controlled locally, in a decentralized manner.

The heterogeneity of the members of a team of robots as well as the information that comes from their various sensors is an advantage that can be used in favor of the success of the mission. For example, in case of an earthquake, a quick picture from a drone can shows the general situation with the possible rescue targets. Then mobile robots like Segway-based ones, more robust and equipped, can be sent to the target area. Meanwhile, some vital elements can be delivered by drones. Or in a coal mine, a first line of robots can obtain images and gas information in order to be used by rescuers. Once the risky environmental information has been obtained, rescuers will be notified, so they can be prepared [14]. If it is safe for rescuers, they can go ahead and reach the robots and then, operate them continuing with the SAR mission.

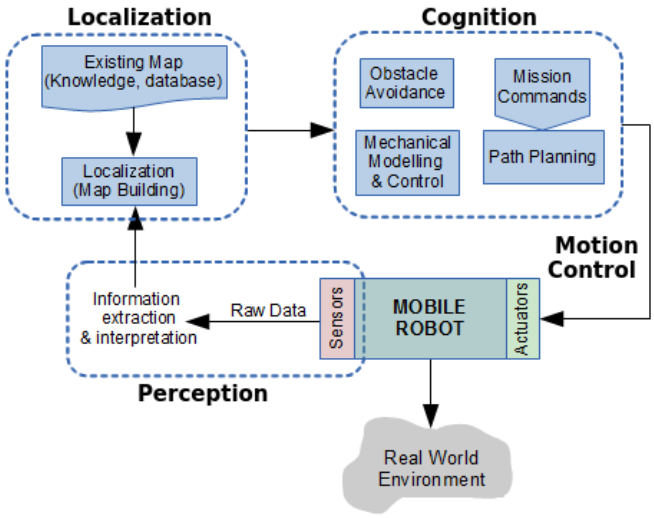


Figure 1. Control structure for autonomous mobile robots

For a collaborative success, the deployed network of robots must use a service software, known as the middle-ware layer.

3.3. Middle-Ware for SAR Robotics

The middle-ware can be seen as the glue that connect everything in the robotic network. It should be designed to allow easy integration of each robot into the network, especially considering that in semi-autonomous robotic implementations for SAR operations, the system configuration time is a critical factor. Another important point that can influence in the direction of the mission is the level of intelligence assigned to the robotic resources for the search process. The middle-ware should provide software interfaces that mask the heterogeneity between the different robots in order to promote cooperation among the robots. This approach allows a more efficient use of the technological resources in order to positively affect the duration of the mission. And considering that time is critical, this will also influence in the final result of the mission.

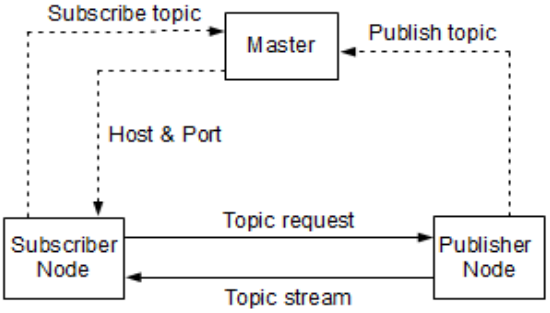


Figure 2. Communication in ROS

The ROS middle-ware consists of nodes, messages, topics and services, and the nodes communicate with each other in a peer-to-peer way (P2P) by publishing messages and subscribing to the

messages posted [15]. An initial event called "naming service" is centralized and based on a master node, which is shown in Figure 2 [15].

A very common and difficult problem when installing a rescue robot in a catastrophe area is the maintenance of a permanent communication. It is expected that communications facilities have been destroyed or inoperative. Therefore, it should be assumed that local communication resources cannot be used, by limiting communications to the native of the system itself [16].

Below is the communication sequence between publisher node and subscriber node:

- Publisher registers a topic (e.g. a laser scan) to the master node informing the point of topic data.
- Subscriber requires the master on how to access to a particular topic.
- Master responds by sending the entry point data, host and part number.
- Subscriber communicates to publisher (host) via TCP or UDP connection requesting for topic data.
- Publisher responds by sending the topic data stream (e.g. laser scan data).

The ROS has been modularly designed. It is organized in packages containing nodes, configuration files, libraries, databases and third party software. So, this packages can be easily integrated into the ROS framework. Exist a wide range of ROS packages available for various robotic implementations such as sensor/actuator drivers, robot path planning and navigation, robot simulation, and others [17].

3.4. Robotic Vision

Computer Vision is a discipline whose main objective is the identification and recognition of objects within digital images, normally acquired by digital cameras. In many applications the identification of moving objects in video captures is required. In addition, in some applications this is required to be achieved in real time. The strategies used for the identification of objects in digital images are based on the analysis of the edges or the study of the whole image or a partial window where the object is probable to be in [18]. The strategies based on edge analysis use techniques to detect lines and edges such as thresholding [19, 20] or by the use of different filters [21] to then extract the characteristics of the shape using digital topology [22, 23] or by recognizing patterns obtained from the border. The strategies based on the analysis of the complete area of the image obtain characteristics of the object contained in the image through training and pattern recognition techniques using neural networks [24], genetic algorithms or using the wavelets transformation [25]. The discretization of the image data, the resolution of the camera, the lack or excess of brightness, the obfuscation, cause lost of clarity or noise in the image, which requires extra treatment [26-28] that in many cases avoid real time requirements.

The robotic vision requires converting the two-dimensional information of the scene obtained from images recorded by cameras into a three-dimensional model in order to recognize objects and places, in order to correctly execute the assigned tasks. Stereoscopy is a technique that imitating human vision, from two images of the same scene achieves a three-dimensional reconstruction [29]. The computation is achieved after the recognition of homologous points in both images. However, this task requires testing the matching of millions of pixels in millions of possible combinations [30] which compromises the requirements for recognition in real time.

The use of commercial 3D cameras for robotic vision can be expensive and heavy. In the other hand, proprietary technology limits several implementations. The aim of the 3D vision research is to develop a cheap, flexible and open solution with stereo cameras that can generate the 3D reconstruction using stereoscopy in different scenarios.

3.5. About the Research Collaboration between Argentina and South Africa

Within the framework of the Scientific-Technological Cooperation Program between the Ministry of Science, Technology and Productive Innovation of the Argentine Republic (initials in Spanish: MINCyT) and the Department of Science and Technology of the Republic of South Africa (DST), the project called "Semi-Autonomous robots for SAR operations" was selected by the Bilateral Commission MINCYT-DST to be executed in the 2014-2016 triennium under code SA / 13/13.

In charge of the South African part was Dr. Glen Bright of the Mechanics Department of the Kwazulu-Natal University, while in charge of the Argentina part was Dr. José Zelasco of the Mechanics Department of the Faculty of Engineering of the University of Buenos Aires.

Phd students Nicol Naidoo and Jorge Kamlofsky have made progress in their researches. Engineering students from both universities have participated in the developments and attended the presentations of the progress of the project.

Throughout the project various equipment was acquired such as: drones, Segways, cameras and others. Committees of both countries have made presentations about the progress of the project and joint publications were achieved [1, 41, 47, 48]

Although it was not possible to present a functional prototype of rescue robot, the development of a middle-ware to integrate several robotic platforms was made. Several stages of a three-dimensional vision system were developed that prove to be very efficient. Much of this is shown in this work.

Challenges of this experience were the technical developments, the interrelation between the researchers of both countries and the active participation of the students.

4. The Vision-Ware: 3D Vision Algorithms and Hardware Integration

Artificial vision technology is based on a resemblance to human natural vision. Following the analogy with human vision, each eye receives a different luminous stimulus, the distance that separates them produces a parallax angle in the observed object. The image obtained in each retina is integrated and reconstructed in the brain that finally generates the perception of the relief. In order to imitate this reconstruction process, given a pair of images (obtained by cameras) called stereoscopic image, it is necessary to solve the following three stages: Calibration, pairing or put in correspondence of homologous points, and 3D reconstruction.

To obtain the three-dimensional reconstruction of an object present in two images, first, it is necessary to find it in both images. It continues with the pairing or finding of homologous points. This is: to find the location of the points of the object in the first image that correspond to the points of the same object found in the second image. Finally, by means of stereoscopy, the 3D coordinates of those points of the object can be calculated.

In this section, a model that for shape recognition uses just a few points of each object is presented. Therefore, it is attractive to use in real time robotic vision. Experimental data is presented.

4.1. Preparing 3D Models: Digital Camera Calibration

To perform the 3D reconstruction of a scene from multiple images by stereoscopy, these must be calibrated. At this point, the Fundamental matrix is a key concept since with the geometric information available, it allows to obtain the epipolar geometry of the scene from uncalibrated images [49].

If a point P of a three-dimensional referential system is projected on the left image as P_L and on the right one as P_R, (as shown in Figure 3) then the image point direction vectors in the same referential system satisfy the equation $P_L.b \times P_R = 0$, where b is the displacement vector between the cameras: $b = (X_c, Y_c, Z_c)^T$, or the coordinates of right camera in the 3D referential system, if the referential system is posed in the left camera. Since the three vectors are coplanar (epipolar geometry), the mixed product gives zero. By the use of the anti-symmetric matrix B:

$$B = \begin{pmatrix} 0 & -Z_c & Y_c \\ Z_c & 0 & -X_c \\ -Y_c & X_c & 0 \end{pmatrix}$$

The mixed product can be expressed as: $P_L \,.\, B \,.\, P_R = 0$. W_L and W_R are the coordinates expressed in pixels in the digital images of the points P_L and P_R expressed in length units related by the calibration matrix C as follows:

$$W = \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha & 0 & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = C.V \qquad (1)$$

or:

$$V = \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{\alpha} & 0 & \frac{-u_0}{\alpha} \\ 0 & \frac{1}{\beta} & \frac{-v_0}{\beta} \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = C^{-1}.W \qquad (2)$$

where $W = (u\,,\, v)^T$ coordinates of the point in images, $(u_0\,,\, v_0)^T$ the principal point of the image, α and β focal distances, $V = (x\,,\, y)^T$ coordinates of the point in length units.

Because the reference is posed on the left camera, also a rotation of the right camera is considered. Matrix R represents the right camera orientation. Thus:

$$W_L.(C^{-1})^T.B.R.C^{-1}.W_R = 0$$

with:

$$F = (C^{-1})^T.B.R.C^{-1}$$

F is the Fundamental matrix [24, 25, 31] with 9 parameters involved: 4 parameters for the camera calibration, 2 for the rotation and 3 for the base b. It is a square matrix of order 3 and rank 2.

When the calibration is known, only the parameters of B and R are unknown. In this case, it is easy to obtain the essential matrix $E = B \,.\, R$ [32] from the Fundamental matrix, since $E = C^T.F.C$ There are several approaches to obtain the 5 involved parameters [33 - 37]. The usual method involves the singular value decomposition of the Essential matrix [38].

Many methods have been proposed in order to calibrate the camera. In [47] is proposed a very simple calibration scheme, which is easy to implement. An assumption made is that it is quite easy to measure the X and Y coordinates of the camera point of view in relation to a 3D referential system when the camera optical axis has a small rotation angle with the Z axis. It is accepted that the error in the Z coordinate of the point of view is less than +/- 1cm. The error in the calibration parameters were obtained and evaluated in [47]. In a later step, assuming that two cameras have a small angle in relation to the Z axis, the calibration parameters error were evaluated and corrected, knowing the distance between two benchmarks in relation to the one calculated by stereoscopy. Regarding the Essential matrix [39], a method for getting the base B and the rotation R, with the solution of linear systems was used [47].

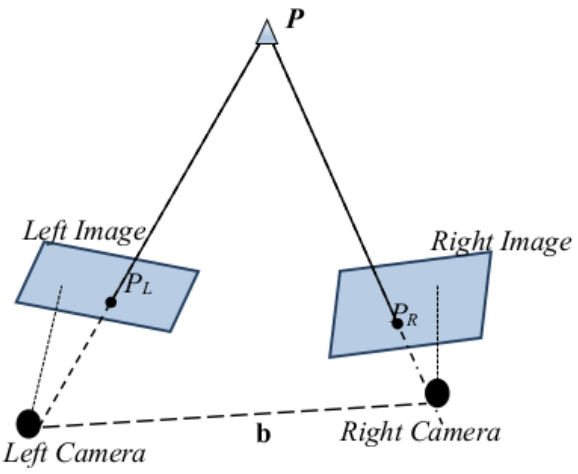


Figure. 3. Epipolar geometry

Calibration parameters are necessary to obtain more precise 3D coordinates of the scene points obtained by stereoscopy.

The interest in recognizing objects within digital images has two main areas of application: obtaining information from images for subsequent human interpretation and processing scenes for autonomous robotic vision [21]. In this last area, for the quick movement and efficient realization of the tasks for which the robot was built, it is especially critical that the processing of the images acquired by the robot, must be done in real time.

The following tasks (detailed below) are necessary to find objects within the images: Image Acquisition, image binarization: separation of objects of interest from the bottom (the complement of objects set), border acquisition: the boundary between objects set and bottom, border simplification: polygonalization, obtaining the curves descriptor pattern, patterns matching. All these tasks have been developed with the premise of realizing as soon as possible, approaching so, to the real time requirements.

Image Acquisition: A Digital Image is a two-dimensional luminous intensity function *f (x , y)*, where *x* and *y* denote the spatial coordinates in the image and the value of *f* at the point *(x , y)* is proportional to the brightness or gray scale of the image at that point.

To acquire a digital image, an image sensor and the capability to digitize the signal from sensor is required. A sensor could be a monochrome or a color camera [21]. The gray scale of images vary depending the image type. RGB mode is commonly used in programming. In the RGB color images each pixel has three components: Red, Green and Blue, where each component represents the amount of each color in values between 0 and 255 (or one byte) where 0 indicates absence of color and 255 means full color. Thus colors can be represented with 3 bytes or 24 bits. Each point of the digital image is called a pixel. A simple color RGB image has millions of pixels. And each pixel can present color values between 0 and 24 bits.

Image Binarization: In Computer Vision it is convenient to work with binarized images. This is achieved by changing the value of the gray scale by a 1 or a 0. The objective is to separate objects of interest from the background of the image by a process known as "segmentation by thresholding". The simplest way to do this consists of traversing each pixels of the image and assigning a in the gray scale "1" if gray scale conditions are higher than a defined threshold value and "0" otherwise. In this way, a "binarization of the image" is achieved [21]: grouping pixels of objects of interest with one value and pixels of the bottom with the other one.

In other cases, thresholds have to limit color ranges. Thus, multiple thresholds have to be defined: maximum and minimum threshold in each channel: red, green and blue.

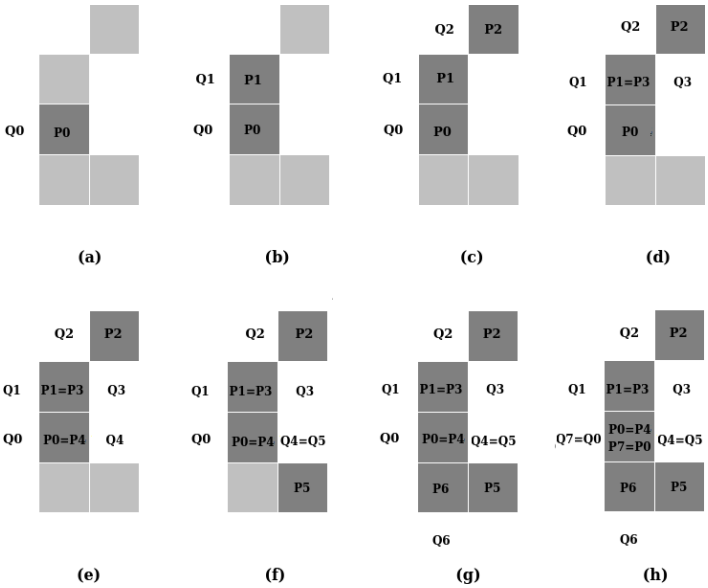


Figure 4. BF8 Algorithm example: (a) Starts in (Q0 , P0). (b) First pair obtained looking 8-neighborhood of P0, going clockwise: (Q1 , P1). Next iterative part illustrated from (c) to (g). Algorithm stops in (h).

The Borders: The objects obtained by binarization of the image are delimited by their borders. A simple algorithm for drawing edges was presented in [22]: BF4 or BF8 depending on whether 4-neighborhoods or 8-neighborhoods are used.

. Figure 4 illustrates algorithm BF8.

BF4 is similar than BF8 but using 4-neighborhoods. With BF4 a thicker edge is obtained, with more points, while with BF8, the edge is thinner, with fewer points. Because of this, the second case is faster.

The edges allow to separate the objects from the background and thus be able to be treated as open sets of the Digital Topology [23]. The shapes of the objects can be characterized from the analysis of their edges.

Border Simplification: By polygonalizing the edges a simplified approximation of the edge curves can be achieved, and thus, an approximate representation of the object hundreds or thousands of times more reduced.

The polygonalization is done iteratively using the convex hull concept [45]: given a set of edge points, those that are inside a convex hull of width ε will be eliminated. In this way, the sets of points that are approximately aligned, will be represented by the first and the last which will form two vertexes of a polygon. It will continue with next group of points and then, a new vertex of the polygon is determined till end point is the first point of the process. Figure 5 shows a diagram describing its operation. More details about this process can be found in [46].

The parameter ε allows to regulate the compromise relationship between quality and performance: a larger ε decreases the quality of the border simplification with greater speed. A small ε ensures resolution in the approach. With a larger ε, an approximation will be obtained more quickly. It is then left to the user to assume the compromise between resolution and performance through this parameter.

Obtaining of the curves descriptor pattern: From the polygonal curve, a descriptor pattern can be obtained based on the discrete evolution of the curvature parameterized by arc length [42]. It consists of a discrete function κ(λ) where κ is the curvature (the interior angle) that corresponds to the evolution of the perimeter of the polygon parameterized by the arc length λ. While the evolution of the perimeter λ varies between [0, 1] because it is parameterized by arc length, the curvature accumulates a final value of 2π because it is a simple curve. It is clear that in the points of the polygon that are not vertexes, the curvature is zero. Therefore, only is interesting its analysis in the vertexes. However, from each vertex of each polygon a descriptor pattern can be initiated that will vary by shifting. After determining the orientation of each object [46], a point is defined from where to start each pattern. Thus, each object will be characterized by exactly one pattern.

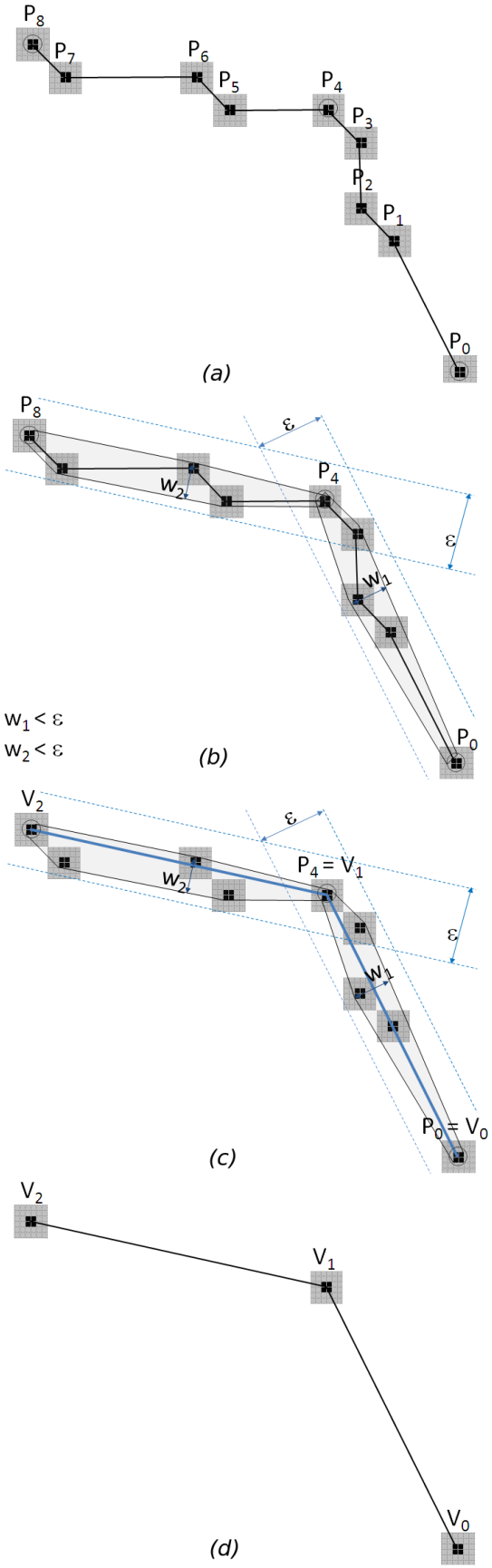


Figure 5: Simplification of curves using the wide convex hull ε. (a) A portion of a digital curve composed of 9 points. (b) The convex hulls of width ε are established. (c) The vertexes of the approximating polygonal are defined. (d) Interior points are eliminated and approximation established.

This pattern is invariant to rotations, translations and scaling, that is convenient when it is necessary to find similar objects instead of exactly equal object.

$$d(\kappa_1, \kappa_2) = \int_0^1 \left(\kappa_1(\lambda) - \kappa_2(\lambda)\right)^2 d\lambda \qquad (3)$$

Figure 6 shows a set of patterns corresponding to different objects. In black is shown similar patterns of similar objects.

Because in images normally objects could be in any orientation, these may be subject to transformations (explained in detail in [43]). Since objects are treated as polygons with reduced number of vertexes, and polygons can be represented just with its vertexes [46], the transformations to all the objects can be performed by a product between the transformation matrix and the matrix formed by the ordered points of the vertexes of this polygon.

4.3. Fast Finding of Homologous Points: Preparing 3D models in Real Time

From the recognition of homologous points in both images (called also: "pairing points in images"), the position in the three-dimensional space can be calculated. However, this task requires the pairing of many points, which compromises performance requirements. A model to finding out homologous points was presented in [40] and mentioned in [1]. In this paper, performance improvements is presented. Experimental data is included.

Description of the Process: When an object is found in both images with the method described above, one homologous point is automatically defined (in both curves): the starting point of both patterns.

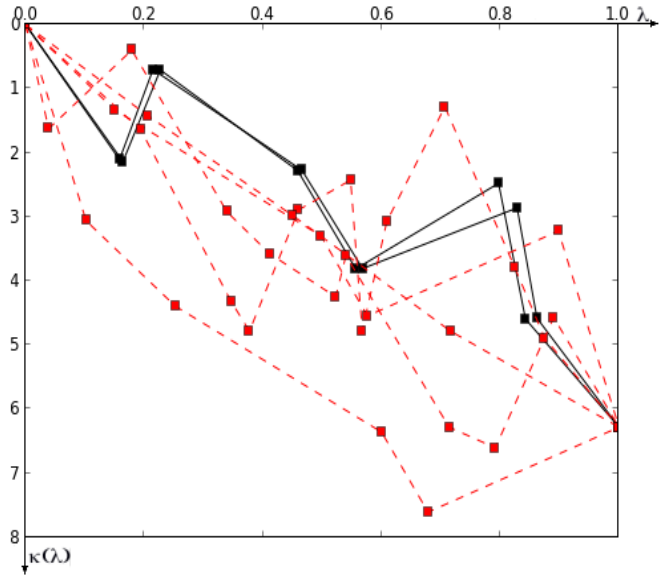


Figure 6: The Curvature Evolution Pattern of various objects.

Thanks to the fact that the curves are parameterized by arc length, and taking into account that the objects are similar, more homologous points can be obtained: this is achieved by traversing the vertexes of one of the curves, its counterpart is sought in the

evolution of the normalized perimeter in the second curve, within a tolerance interval. For example, if a tolerance of 5% is performed and for a vertex *i* of the pattern of the first object with coordinates $(\lambda_{1i}, \kappa_{1i}) = (0.200, 1.000)$, a vertex *j* in the pattern of the second object is wanted: $(\lambda_{2i}, \kappa_{2i}) / 0.250 > \lambda_{2j} > 0.150$ and $1.050 > \kappa_{2j} > 0.950$. In case of a correspondence, the pair of vertexes (*i* , *j*) is added to the set of homologous points.

Performance Improvements: In [1] and in [40] it was shown that for the finding of the homologous points of an object in two images, it took 4.9 seconds. However, contemplating certain restrictions, the same finding could be achieved in times less than 1 second, for the same images.

In the process of objects recognition within digital images, it is intended to improve the performance in the reading of the images (mostly binarized or filtered) that the process provides as parameters.

This operation can be done faster without losing important details under the following assumptions:

- No objects of interest less than 10px wide.
- In a binarized image (black and white), the R, G and B components of the pixel are similar: 0 or 1.
- During the tracking of a moving object, the acceleration and velocity of movement of the object are known.

With these assumptions:

- The horizontal direction of reading of the image is done in discrete 10px jumps instead of being traversed pixel by pixel. If a value of interest is found in one of these jumps, 10 positions are retracted, and the current position is returned by consulting the values of the involved pixels one at a time and extending to the next 10 positions. If this operation was performed in the previous jump, the survey of the previous 10 positions is not carried out.
- If the image is binarized, the R, G and B components of any pixel are similar. Then, instead of requesting the R, G and B components, only one of them is consulted.
- If the speed and acceleration of the object is known, its location can be estimated, so that its search can be done within a reduced window, and not in the whole image.

As a consequence of the above, a great reduction in the time of the reading process of the images is obtained.

4.4. Stereoscopy: Building the 3D Model

Starting with two images, knowing homologous points in both images, by the use of stereoscopy it is possible to perform a 3D re-composition of a scene based on these points. The more homologous points are known, the better the reconstruction will be. Thus, calculation of the 3D coordinates of each of these point (generically named here as $\boldsymbol{M_q}$) is easily obtained by contributions of information from both images [30]. The spatial coordinates of the generic point $\boldsymbol{M_q = (x , y , z)^T}$ are obtained by using the scene reconstruction equation (Equation 4) applied to each image:

$$\begin{pmatrix} \check{\imath}.R - U.\check{k}.R \\ \check{\jmath}.R - V.\check{k}.R \end{pmatrix}.M_q = \begin{pmatrix} \check{\imath}.L - U.\check{k}.L \\ \check{\jmath}.L - V.\check{k}.L \end{pmatrix} \quad (4)$$

Where $\boldsymbol{R}$ is the rotation matrix of the camera, $\boldsymbol{L = R\,.\,S}$ with $\boldsymbol{S}$ the position of the camera in the general referential system, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are *x* and *y* coordinates in length units of the projection of point $\boldsymbol{M_q}$ in each image. It is important to note that in this equations some parameters calculated in the Calibration step are used. Since $\boldsymbol{M_q}$ is unknown then $\boldsymbol{M_q}$ can be called as $\boldsymbol{X}$, and (4) can be stated briefly as: $\boldsymbol{A\,.\,X = B}$ where $\boldsymbol{A}$ is a 2 x 3 matrix, and $\boldsymbol{B}$ is a 2 x 1 matrix. Then, contribution of the left image is: $\boldsymbol{A_L.\,X = B_L}$ and the contribution of the right image is: $\boldsymbol{A_R.\,X = B_R}$.

Joining contributions of both images:

$$\begin{pmatrix} A_L \\ A_R \end{pmatrix}.X = \begin{pmatrix} B_L \\ B_R \end{pmatrix} \quad (5)$$

Naming $\boldsymbol{A = (A_L\,,\,A_R)^T}$ and $\boldsymbol{B = (B_L\,,\,B_R)^T}$, (5) can be presented as:

$$A.X = B \quad (6)$$

Where $\boldsymbol{A}$ is a 4 x 3 matrix, and $\boldsymbol{B}$ is a 4 x 1 matrix. Pre-multiplying by $\boldsymbol{A^T}$ in both members of (6), and then by $\boldsymbol{(A^T\,.\,A)^{-1}}$, 3D coordinates of the re-construction can be calculated as stated in (7):

$$M_q = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = (A^T.A)^{-1}.A^T.B) \quad (7)$$

The system is not canonical, that is, the independent term $\boldsymbol{B}$ does not correspond exclusively to the values measured in the images, therefore, even if the error of those measurements was known, the equations could not be weighted and the variance of the elements of the unknown vector could not be calculated.

Once points in spatial coordinates are obtained, the three-dimensional model of the scene can be achieved.



Figure 7: A notebook connected to two regular cameras mounted in a platform with known distance and orientation.

The three dimensional vision system is constructed using two regular cameras. One camera is placed in the origin of the referential system. The other one is located on the *x* axis, and thus, the referential system is fixed. The distance between both cameras is known, as well as the orientation angles. The greater the distance, the more accurate are the obtained spatial coordinates by stereoscopy. Current cameras can be connected directly to USB ports. And more than one camera can be connected to regular computers, as shown in Figure 7. Cameras are controlled by the vision software. Normally, drivers of regular cameras exist in most

operative system; if not, drivers have to be installed from manufacturer's site.

The Segway based robot is built to work in harmful environment. Thus, it is recommended to install an industrial computer, which is designed to work under more difficult conditions than regular computers.

Also with small size computers like Raspberry Pi-3 Model-B, 3D vision systems can be built. It is a good option in cases when weight and size availability is limited. In these boards with less than 50 grams and 50cm^2, 64-bit processor, wireless connectivity and 4 usb ports are included. The base operative system is Raspbian: a specific Linux distribution. If it is not enough, other Linux distributions or Windows 10 for IOT could be installed. These boards can be used in a drone based robot. A different robotic vision system can be constructed with just one camera: images are taken after moving a known distance. The greater the distance between each image, the more accurate are the obtained 3D coordinates. Therefore, precision of 3D re-construction can be controlled by varying the distance between photos, with the height known. Figure 8 shows a Phantom 3 drone with an HD camera.



8: Phantom 3 drone camera with a stabilization mechanism.

In this project is tested the use of regular, IR and thermal cameras.

4.6. The Experiment

Brief Description: The experiment consists of obtaining the homologous points of an object present in two images, and measuring of CPU time consumed in each part of the process repeatedly.

Equipment Used and Settings: For the acquisition of images, two Logitech cameras C170 230mm distant were used. Being conventional cameras, the segmentation is achieved by placing objects on a white background. The algorithm was programmed in Python 2.7. The computer used is an HP Pavillion notebook with AMD A10 processor with 12Mb of Ram and 4 cores. The operating system used is Kali Linux: a Linux distribution based on a Debian kernel. Each image acquired has 640 x 480 pixels.

Values used: The width of convex hulls was ε = 4. Tolerance in finding homologous point was 5%.

Results Obtained: The set *S* of homologous points obtained is:

S = {[(390 , 242) , (187 , 191)] , [(417 , 248) , (212 , 198)] , [(458 , 272) , (243 , 216)] , [(514 , 273) , (310 , 213)] , [(511 , 354), (307, 295)] , [(440 , 347) , (233 , 292)] , [(436 , 333) , (225 , 277)], [(351 , 329) , (146 , 281)] , [(274 , 294) , (73 , 256)] , [(250 , 265), (46 , 221)] , [(253 , 257) , (49 , 215)] , [(292 , 274) , (94 , 230)], [(341 , 277) , (140 , 227)] , [(388 , 244) , (181 , 196)]}

Between brackets the homologous points are presented. The first pair corresponds to the first image, while the second pair has its homologous in the second image. Points are expressed in pixels.

In Table, repeated measurements of CPU time consumed in each part of the process, is shown.

From the data showed in Table the average of the time of the process is 0,9199 seconds (less than 1 second). Figure 9 shows images used for the experiment with the homologous points.

Table 1: CPU times (in seconds) of pairing points in images

# test	Model Image (Img 1)			Target Image (Img 2)			Matching	Pairing	Total
	Segment. matrix	Borders	Patterns	Segment. matrix	Borders	Patterns			
1	0,2310	0,0626	0,1121	0,2257	0,0663	0,1015	0,0253	0,0261	0,8506
2	0,2247	0,0578	0,1224	0,2313	0,0589	0,1027	0,0277	0,0237	0,8492
3	0,2435	0,0682	0,1632	0,2356	0,0708	0,1592	0,0262	0,0232	0,9899
4	0,2339	0,0684	0,1689	0,2309	0,0663	0,1632	0,0237	0,0266	0,9819
5	0,2233	0,0725	0,1662	0,2254	0,0668	0,1568	0,0263	0,0260	0,9633
6	0,2254	0,0717	0,1115	0,2553	0,0686	0,1546	0,0309	0,0276	0,9456
7	0,2353	0,0627	0,1129	0,2349	0,0634	0,1085	0,0262	0,0239	0,8678
8	0,2328	0,0656	0,1191	0,2283	0,0674	0,1093	0,0264	0,0226	0,8715
9	0,2286	0,0632	0,1604	0,2397	0,0677	0,1620	0,0200	0,0236	0,9652
10	0,2379	0,0632	0,1139	0,2216	0,0692	0,1612	0,0225	0,0245	0,9140

Previously, the relationship between the level of autonomy and remotely assisted control in semi-autonomous robots for SAR tasks was discussed, and this is presented in this section. A hybrid control architecture is defined: an internal control block for automatic tasks, and a remotely assisted management and control scheme.

The robots have an internal control block that allow efficient time running of automatic tasks as: sensing, moving, obstacle avoidance, local path planning, etc. The robotic system have also a higher level of control called the "Navigation Control Center" (NCC) as shown in Figure 10. The objective of the NCC is to record and update the mapping models with the data acquired from the field and to assign a coverage area to robots and them to determine its navigation algorithm.

The control panel is the part of the NCC (not shown in the NCC block in Figure 10) where robots can be controlled remotely. The operators can request for information or tasks interacting through the graphical user interface (GUI) of the control panel, that will be developed in order to facilitate remote operation. A request made in the NCC is a high level task meanwhile in the local control block of the robot, it is transformed into a specific sequence of tasks just programmed. NCC and local control blocks are connected by a radio link.

The robots of the network will be powered by the ROS middleware. It consists of a master - subscriber/publisher nodes as explained in Section III and shown in Figure 2. ROS consist of the following software modules:

- Perception: ROS software drivers consist of packages of software that connect physical sensors with the operative

system. Normally drivers are installed in the ROS in its last version. But in some specific cases drivers have to be installed. Depending of the kind of robot, different sensors will be mounted. While in a basic drone just one camera is carried, in the mobile robot based on a Segway platform a plenty of sensors could be installed: cameras, microphones, temperature, humidity, radiation, gas, etc.

- Localization: It is the module of software that calculates the position of the robot in the field, and dynamically builds the maps of newest areas or updates maps of known areas. The new information is shared with the NCC to optimize SAR process.

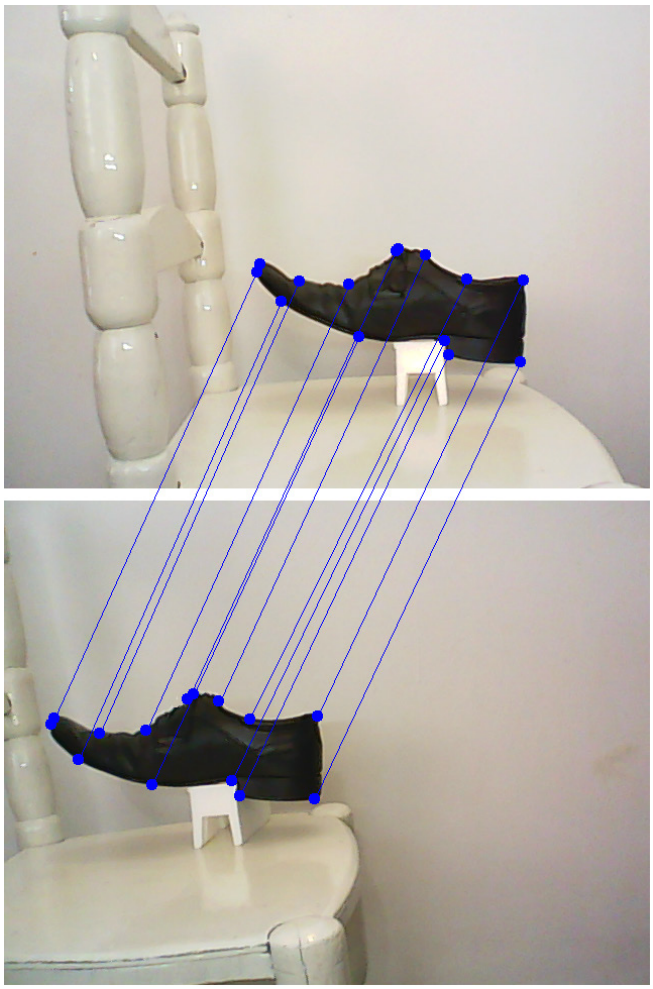


Figure 9: Homologous points of an object in two images.

- Cognition: This module has three important functionalities. The first one is the motors control module: provides electrical control of the motors that make physical actions of the robot such as to catch an object or to move from one site to another. The second functionality is obstacle avoidance: considers safety measures to prevent harm to persons, damages to the environment and to the robot itself. The last one is the local path planning: the objective is to make a surveillance in an area searching for humans that may require help or needs being rescue.
- 3D vision algorithm: this involves the processing of data from images acquired by the stereo cameras. By the 3D vision algorithm, data (images) are transformed into a higher level of information like an obstacle to be avoided or a possible alive person. This information must be shared with the NCC (so rescue personnel can be sent).

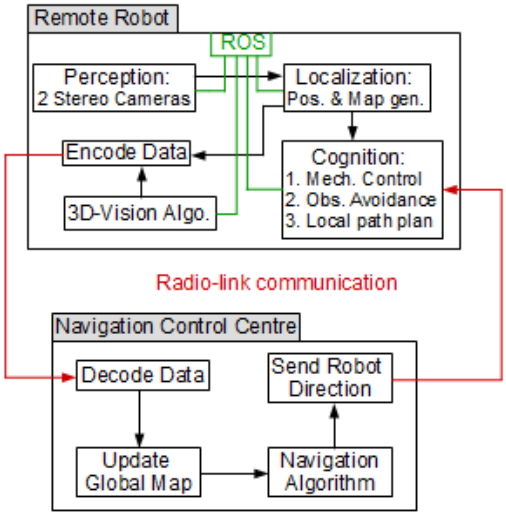


Figure 10. Robot communication and control architecture

A design factor that is a non-trivial problem is the possibility of radio-connectivity problems due to lost of signal strengths caused by interferences of the structures and constructions or because the geography of the site. To solve it can involve an in-depth research analysis, as discussed in [44].

6. Conclusions

In this paper, the need and convenience of the development for its later use in SAR missions of various semi-autonomous robots was introduced and presented.

Collaboration for research between Argentina and South Africa was established, which allowed researchers to make contributions for the design, development, simulations and testing of semi-autonomous robots based on different physical platforms: Segway and drone-based. A hybrid control architecture was presented between the remote robots and the control and navigation center. Interesting improvements in the 3D vision algorithms were presented, which seem to be promising since they would contribute to achieve three-dimensional modeling in real time.

Research results achieved during the binational project related to developments, knowledge exchange and active participation of students from both countries were presented.

Acknowledgement

The research presented in this paper is part of a joint collaboration between the University of KwaZulu-Natal, in South Africa, and the University of Buenos Aires, in Argentina. The

authors wish to thank their respective governments for funding this research and for making the collaboration possible.

References

[1] N. Naidoo, G. Bright, R. Stopforth, J. Zelasco, F. Ansaldo, M. Bergamini, and J. Kamlofsky, "Semi-autonomous robot control system and with 3D vision scheme for search and rescue missions: A joint research collaboration between South Africa and Argentina." In Mechatronics and Machine Vision in Practice (M2VIP), 24th International Conference on (pp. 1-6). IEEE, 2017.

[2] G. De Cubber et al., "Search and rescue robots developed by the European Icarus project, 7th Int." In Workshop on Robotics for Risky Environments. 2013.

[3] Greer D., P. McKerrow, and J. Abrantes, "Robots in Urban Search and Rescue Operations", © ARAA, In 2002 Australasian Conference on Robotics and Automation, 2002.

[4] B. Doroodgar et al., "The search for survivors: Cooperative human-robot interaction in search and rescue environments using semi-autonomous robots." In Robotics and Automation (ICRA), 2010 IEEE International Conference on. IEEE, 2010.

[5] J. Casper and R. Murphy, "Human-Robot Interaction during the Robot-Assisted Urban Search and Rescue Response at the World Trade Center", IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol. 33, No. 3, 2003.

[6] R. Murphy and S. Stover, "Rescue Robots for Mudslides: A descriptive study of the 2005 La Conchita Mudslide Response", International Journal of Field Robotics, Wiley, 2008.

[7] K. Kleiner, "Better robots could help save disaster victims", 2006.

[8] S. Brenner, S. Gelfert and H. Rust. "New Approach in 3D Mapping and Localization for Search and Rescue Missions." CERC2017: 105-111, 2017.

[9] P. Fritsche. and B. Wagner, "Scanning Techniques with Low Bandwidth Radar for Robotic Mapping and Localization," in Informatics in Control, Automation and Robotics (INCINCO 2015). IEEE: 321-335, 2015.

[10] A. Kumar et al., "Search and rescue operations using robotic darwinian particle swarm optimization." In Advances in Computing, Communications and Informatics (ICACCI), 2017 International Conference on. IEEE, 2017.

[11] C. Mouradian, S. Yangui, and R. Glitho. "Robots as-a-service in cloud computing: search and rescue in large-scale disasters case study." In 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2018.

[12] A. Elfes, "Using occupancy grids for mobile robot perception and navigation." Computer 6: 46-57, 1989.

[13] R. Siegwart and I. Nourbakhsh, "Introduction to Autonomous Mobile Robots", MIT Press, Cambridge, Massachusetts, 2004.

[14] J. Zhao et al. "A search-and-rescue robot system for remotely sensing the underground coal mine environment." Sensors 17.10, 2017.

[15] N. Naidoo, G. Bright, and R. Stopforth, "A Cooperative Mobile Robot Network in ROS for Advanced Manufacturing Environments", in Proceedings of the International Conference on Competitive Manufacturing (COMA'16), 2015.

[16] S. Shin et al. "Communication system of a segmented rescue robot utilizing socket programming and ROS." Ubiquitous Robots and Ambient Intelligence (URAI), In 14th International Conference on. IEEE, 2017.

[17] List of ROS packages for Indigo, http://www.ros.org/browse/list.php, last accessed 18 July 2017.

[18] D. Zhang, G. Lu, "Review of shape representation and description techniques". Patt. Recogn., 37 (1),1-19, 2004.

[19] L. Davis, "A survey of edge detection techniques."Computer graphics and image processing 4.3: 248-270, 1975.

[20] A. Rosenfeld and A. Kak. "Digital picture processing". Vol. 1. Elsevier, 2014.

[21] R. Gonzalez, R. Woods. "Digital image processing."Addison-Wesley Publishing Company, 1993.

[22] A. Rozenfeld, "Digital Topology", The American Mathematical Monthly, 86(8) pp. 621-630, 1979.

[23] U. Eckhardt and L. Latecki. Digital topology. Inst. für Angewandte Mathematik, 1994.

[24] H. Rowley, S. Baluja and T. Kanade. "Neural network-based face detection."IEEE Transactions on pattern analysis and machine intelligence 20.1: 23-38, 1998.

[25] A. Fernandez Sarría, "Estudio de técnicas basadas en la transformada Wavelet y optimización de sus parámetros para la clasificación por texturas de imágenes digitales". Diss. Universitat Politècnica de València, 2007.

[26] J. Weaver, et al. "Filtering noise from images with wavelet transforms."Magnetic Resonance in Medicine 21.2: 288-295, 1991.

[27] L. Rudin, S. Osher, and E. Fatemi. "Nonlinear total variation based noise removal algorithms."Physica D: Nonlinear Phenomena 60.1: 259-268, 1992.

[28] T. Nguyen, I. Debled-Rennesson. "Curvature Estimation in Noisy Curves." In: W. Kropatsch, M. Kampel, A. Hanbury (eds.) Computer Analysis of Images and Patterns, LNCS, vol. 4673, pp 474-481, Springer, Heidelberg, 2007.

[29] H. Moravec. "Robot spatial perceptionby stereoscopic vision and 3d evidence grids."Perception, 1996.

[30] J. Zelasco et al. "Computer vision in AUVs: automatic roto-rectification of stereo images."OCEANS 2000 MTS/IEEE Conference and Exhibition. Vol. 3. IEEE, 2000.

[31] Q. Luong and O. Faugeras, "Determining the Fundamental matrix with planes:Instability and new algorithms", Proc. Conf. on Computer Vision and Pattern Recognition, pp 489-494, 1993.

[32] Q. Luong and O. Faugeras,"Self-Calibration of a moving camera from Point correspondences and fundamental matrices", International Journal of Computer Vision,22 (3), pp 261–289, 1997.

[33] H. Longuet-Higgins, "A Computer Algorithm for Reconstructing a Scene from Two Projections", Nature, 293 (10), pp 133-135, 1981.

[34] J. Heikkila, "Geometric camera calibration using circular control points", IEEE Transactions on Pattern Analysis and Machine Intelligence22(10), pp. 1066-1077, 2000.

[35] Z. Zhang. "A flexible new technique for camera calibration",IEEE Transactions on Pattern Analysis and Machine Intelligence22(11), pp. 1330-1334, 2000.

[36] Z. Zhang,"Camera calibration with one-dimensional objects"IEEE Transactions on Pattern Analysis and Machine Intelligence26(7), pp. 892-899, 2004.

[37] H. Stewénius, C. Engels and D. Nister, "Recent Developments on Direct Relative Orientation", ISPRS Journal of Photogrammetry and Remote Sensing 60, pp 284-294, 2006.

[38] R. Hartley and A. Zisserman, "Multiple View Geometry in Computer Vision", Cambridge University Press, 2003.

[39] M. Kalantary and F. Jung, "Estimation Automatique de l'Orientation Relative en Imagerie Terrestre.", XYZ-AFT, 114, pp 27-31, 2008.

[40] J. Kamlofsky and M. Bergamini, "Rápida Obtención de Puntos Homólogos para Vision 3D". In VI Congreso de Matemática Aplicada, Computacional e Industrial (VI MACI 2017) , 2017.

[41] M. Bergamini, F. Ansaldo, G. Bright, and J. Zelasco. "Fundamental Matrix: Digital Camera calibration and Essential Matrix parameters". In International Journal of Signal Processing: 120-126, 1 (2016).

[42] T. Kanade and O. Masatoshi. "A stereo matching algorithm with an adaptive window: Theory and experiment."IEEE transactions on pattern analysis and machine intelligence 16.9: 920-932, 1994.

[43] E. Lengyel. "Matemáticas para Videojuegos en 3D". Second Edition, Cengage Learning, 2011.

[44] S. Chouhan, D. Pandey and Y. Chul Ho, "CINeMA: Cooperative Intelligent Network Management Architecture for Multi-Robot Rescue System in Disaster Areas", in Proceedings of the International Conference on Electrical, Electronics, Computer Science, and Mathematics Physical Education and Management: 51-61, 2014.

[45] M. De Berg et al. "Computational Geometry", ISBN 3-540-61270-X Springer-Verlag Berlin Heidelberg New York, 1997.

[46] J. Kamlofsky and M. Bergamini. "Patrón de Evolución Discreta de Curvatura y Concavidad para Reconocimiento de Formas." CONAIISI, 2013.

[47] M. Bergamini, F. Ansaldo, G. Bright and J. Zelasco.. "Fundamental Matrix: Digital camera calibration and Essential Matrix parameters". In 16th International Conference on Signal Processing, Computational Geometry and Artificial Vision 2016 proceedings (ISCGAV), 2016.

[48] N. Naidoo et al. "Optimizing search and rescue missions through a cooperative mobile robot network." Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), IEEE, 2015.

[49] Q. Luong et al. On determining the fundamental matrix: Analysis of different methods and experimental results. INRIA, 1992.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 4, 354-362 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

MARWIN: Localization of an Inspection Robot in a Radiation-exposed Environment

Andre Dehne*, Nantwin Möller, Thorsten Hermes

Andre Dehne, hochschule 21 gemeinnützige GmbH, 21614 Buxtehude, Germany

ARTICLEINFO

Article history:
Received: 21 June, 2018
Accepted: 14 August, 2018
Online: 28 August, 2018

Keywords:
Mobile robots
Localization
Radiation detection

ABSTRACT

MARWIN is a mobile autonomous robot platform designed to carry out maintenance and inspection tasks in the European XFEL. The XFEL is an accelerator plant which is operated in Hamburg, Germany. The robot system consists of an four-wheel drive chassis and a scissor lift for easy inspection and maintenance tasks. Through this manipulator and the chassis, the robot system acquires three degrees of freedom. MARWIN is intended for autonomous radiation measurements along the XFEL research facility and thus needs accurate localization. The facility describes a straight tunnel and consists partly of irregular structures and also of sections with almost no obstacles. In the 1000 meter long sections in which MARWIN operates, the robot must approach the facilities to a few centimeters, but must not touch them. For this purpose, different localization methods were tested and checked for accuracy. Furthermore, the influence of radiation on the localization is investigated.

1 Introduction

This paper is an extension of work originally presented in International Conference on Research and Education in Mechatronics (REM) 2017 [1]. Within the research co-operation of the "hochschule 21" and the German Electron Synchrotron (DESY) of the Helmholtz Association a "mobile and autonomous robot for maintenance and inspection" (MARWIN) was developed. For targeted radiation measurement the robot navigates along the new accelerator of the research facility European XFEL. Therefore different systems like localization, drive, charge and lifting are required. After the two-year project, which had to proof that the radiation measurements in accelerator systems can be reliably executed by a robot, which is build of hardware that comes from the consumer sector and is not protected from the radiation, a second robot was build. While the testing phase weaknesses were detected so that improvements could be implemented in the second robot.

1.1 Motivation

The accelerator of the research facility European XFEL is world wide the only one generating 27000 X-ray flashes a second[1]. Compared to other research facilities the XFEL is generating 225 times more X-Ray flashes per second. This improvement grants the enforcement of certain experiments. This results in a high demand for that technology. Therefore it is important to maximize the run-time of the accelerator and thus the effectively usable time for research purposes and experimentations. In return, this means that minimizing the shutdown-time for maintenance is necessary. An automated inspection can help to reduce this time. Through information gathered in advance about certain conditions in the system, maintenance work can be carried out in a targeted manner. In addition, the system must be cleared before it can be entered after a shutdown. This is important to ensures that the radiation generated during operation already is decayed and the accelerator tunnel is safe to enter[2]. An automation of this task, which is otherwise performed manually by employees of the radiation protection department, reduces the burden of these persons and leads to an accurate measurement with a high repeatability.

*Andre Dehne, hochschule 21 gemeinnützige GmbH, Harburger Str. 6, 21614 Buxtehude, Germany, dehne@hs21.de
[1] see https://www.xfel.eu/facility/comparison/index_eng.html
[2] see https://www.xfel.eu/facility/safety_and_environment/index_eng.html

https://dx.doi.org/10.25046/aj030436

1.2 Conditions

A robot system is to be developed that can carry out inspections during accelerator operation. Thus, the availability of the system but also the efficiency of troubleshooting and diagnostics, maintenance and repairs can be increased. The robot should be equipped with a manipulator due to the sometimes tight and difficult to access spatial conditions. This is needed to take measurements on the various components. In addition, a drive system is to be used which enables collision-free movement in the tunnel. The robots are used in the 3.2 *km* long European XFEL tunnel. This is divided into several sections by interlock doors. The robots should measure the radiation in the different sections as autonomously as possible. The current status of the robot systems should always be visible from the monitoring center. In addition, manual intervention by the monitoring center must be possible. The robots will perform two deployment scenarios.

- Scenario 1: For autonomous driving, measuring positions and other data are transmitted by the monitoring center. Based on this configuration, the robot travels the predetermined distance along the accelerator and selectively records radiation measurement values.
- Scenario 2: Manually controlled, the monitoring center can remotely drive the robot to a certain measuring position in order to carry out punctual measurements on the accelerator. In addition, in the unlikely event of autonomous driving failure, the robot is still mobile.

The monitoring center must always be able to switch between scenarios via remote access. The measured radiation values are recorded by a measuring device and then processed by servers outside of the accelerator facility. The individual measurements are linked with the current position data by the robot system.

2 Related work

The application areas of automation and robotics are steadily increasing and more and more solutions are being developed. Nowadays robots can be found in the most diverse areas. For example, at home in the garden for lawn mowing, in the house for vacuuming, in the industry for manufacturing, as transport systems in warehouse logistics or also for support in home care [2, 3, 4].

Another area of responsibility for robots lies in the so-called 4D environments. These environments are characterized by the extreme conditions: Dirty, dull, dangerous or distant conditions. The fourth D is also mentioned for dear conditions where saving money in expensive projects can be achieved through robots[3].

More and more industrial companies prefer the robotics to manual labor, for example in hazardous work environments or in dull tasks [5, 6]. Especially when it comes to keeping people from dangers, developments have been driven forward. They help, for example, with bomb investigations or space exploration [7]. The TRADR project[4] and the Curiosity mission[5] provide important insights in this regard. Another danger arises when dealing with radioactive substances. Here too, first attempts have already been made to use robots in environments with nuclear hazards.

In particle research with accelerator systems, radioactivity also arises. The Large Hadron Collider (LHC) in Geneva also works on mobile solutions to keep dangerous radiation away from humans [8]. At the LHC a rail-based robotic system was installed to perform inspection work in this hazardous environment. It is equipped with various sensors to record local conditions. Since 2016, TIM (Train Inspection Monorail) has been driving through the tunnel hanging on a rail system[6].

Rail-based systems are inherently rigid and can only be used to a limited extent for later additions. The use of freely moving robots in accelerator systems with extended possibilities of use has not yet been sufficiently researched. Localization is an important part of this. It is unclear how a robot must be designed so that it can reliably work freely in such environments.

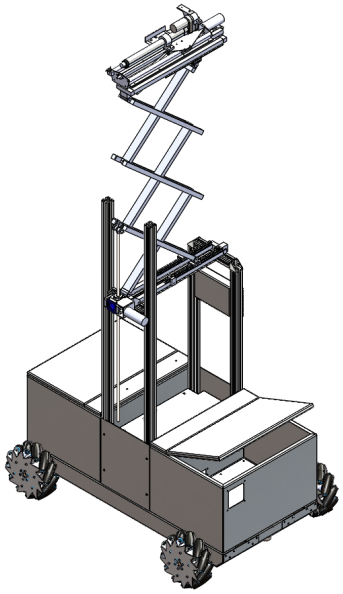


Figure 1: Schematic illustration with mecanum wheel drive in the chassis and scissor lift on top.

[3] see https://www.bernardmarr.com/default.asp?contentID=1195
[4] see http://www.tradr-project.eu
[5] see https://mars.nasa.gov/msl/
[6] see https://home.cern/about/updates/2016/11/meet-tim-lhc-tunnels-robot

3 System overview

Figure 1 shows the developed prototype of a robot system for use in the European XFEL. The robot consists of a mobile drive unit in the lower area and an attached scissor lift for special measuring and inspection tasks.

The drive system of the robot was realized with the mecanum technology [9]. Equipped with four Mecanum wheels, an omnidirectional mobility is achieved. The omnidirectional freedom of movement allows the robot to move better in narrower sections of the XFEL tunnel.

The special feature of a mecanum wheel are the 45 degree angle mounted freewheel rollers, which represent the running surface of the wheel. The size and load capacity of the wheel can be selected according to requirements. There are two different Mecanum wheels. These differ in the arrangement of the freewheel rollers. For the drive with four wheels, two wheels of each type are needed. During assembly, it is important that the pivot of the freewheel rollers points towards the vehicle center.

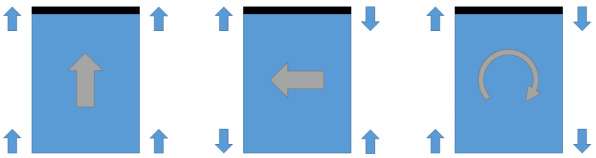


Figure 2: Three examples for the mechanical drive (left: forward; middle: sideways; right: turn on the spot). The arrows indicate the direction of rotation of the wheels.

To illustrate the operation of the mecanum drive, figure 2 shows three different motion scenarios. Driving the wheels in different variations gives corresponding movement patterns. These are enabled by the freewheel rollers. Depending on the direction of rotation of the wheels, movement forces are generated. These can be overlaid or partially canceled out. The resultant force gives the direction of movement. Since there is not much room in the tunnel for turning maneuvers or similar, this property is highly important to avoid possible damage.

For measuring the radiation, as already mentioned, a measuring probe is carried along. This should be guided along the linear accelerator during various measuring runs. There are different scenarios here. One of them is to run along horizontally at a certain height. There are also open structures where the height can vary. For measurement along the accelerator, the probe must be variable in height and positioning. In addition, the robot must not exceed a height of 1.70 m to take his parking position in the charging station.

To meet these requirements, a scissor lift was constructed.With the scissor lift it is possible to reach a big stroke although the basic height is low. For this purpose, three pairs of scissors are used (see figure 1). In order to move the probe horizontally, a platform with a shifting mechanism was set up on top of the scissor lift. This moves the mounted linear actuator up to 30 cm to the outside. The measuring probe is attached to the linear actuator and can thus be moved horizontally by a further 30 cm. To enable the measurement on the irregular structures, a tilt mechanism has been installed. This tilts the platform, mounted on the scissor lift, by up to 30 degrees. The mechanisms are driven by DC gearmotors. For the vertical stroke a spindle-type lifting gear is used to convert the rotary motion of the motor into a lifting movement of the scissor lift. The motor of the displacement mechanism drives a spindle on which a carriage is moved back and forth, depending on the direction of rotation. To move the tilting mechanism, a linear actuator is used.

The robot is a mobile electronic system that is battery powered. Therefore, several charging stations are mounted in the tunnel to charge the battery. Since the robot performs its tasks autonomously, the charging process must also be carried out autonomously. The safety aspect has to be taken into special consideration. Therefore the charging station may only be active when the robot is in the charging station.

In order to ensure safe charging, the two internal contacts of the charging station must be actuated in order to release the charging current at the external charging contacts. For the precise positioning of the robot in the charging station, guide rails were installed on the right and left. These center the robot when driving into the charging station. This ensures that the contacts always lie directly on top of each other.

3.1 Power unit

The power supply of the robot system is realized with a lithium iron phosphate battery (LiFePo4) from the manufacturer ”Super B”. The model ”SB12V100E-ZC” was used[7]. This battery is protected by an external undervoltage protection and was selected due to its high rated capacity of 100 Ah at a nominal voltage of 13.2 V (1320 Wh). The large capacity in relation to the small size and the low weight of the battery is also crucial for the selection of the battery.

The measured current consumption of the entire system is between 15 to 18 A. Thus, the system can be powered up to five hours. That means, that the system can be operated long enough to carry out measuring trips and to get back to a charging station.

3.2 Hardware

The installed IT components are divided as follows:

- Two main computers,
- two single-board computers of the ”Odroid XU4” type with eMMC flash memory,
- one Raspberry Pi 3,

[7] see http://www.super-b.com/en/aviation/utility-avionics-batteries/sb12v100e-zc

- one router,
- one network switch,
- two control units
- six motor driver and
- several sensors and cameras.

The control of the entire system is based on the two main computers, one of which controls the processes and the other idles. The single-board computers are connected to the actuators and sensors. They are used as an interface to the main computers, process sensor signals, control the actuators and output status messages. The control of the lifting system is realized with the Raspberry Pi 3. Router and network switch are required for internal and external communication. With the sensors and cameras, the environment of the robot is observed.

3.3 Sensors

For the localization there are two 2D laser scanners of the type "UST-10LX"[8] manufacturer Hokuyo installed. One in front and one in the back of the robot. Within a range of 270 degree and up to a distance of 10 m, the laser scanner records the surroundings in a single line. This data is also used for the orientation of the robot to the tunnel wall and to prevent collisions. If an object is in the route and is less than 2 m away, a warning is issued. If the object is less than 1 m away, the robot stops to cause no damage. In this case, the monitoring center can intervene manually. For this purpose, eight cameras are installed, which allows complete vision all around the robot. In addition to the camera systems for remote monitoring there is a CCD camera for detecting QR codes, which are installed permanently in the XFEL tunnel. The QR codes contain information about the current location and thus serve the absolute localization of the robot system.

In order to monitor the movement speed of the robot, odometry sensors are used. Hall effect sensors are installed, which incrementally record the rotational movement of the motors. This actual value is processed by a controller to adjust the wheel speed. The regulation of the lifting system is carried out according to the same principle.

3.4 Redundance

With a redundant system a single point of failure should be excluded. Therefore, components such as the main computer are built in duplicate as a master and slave combination. In the event of a fault, the slave can take over the task of the master and restart it. Thus, a total system failure can be avoided and important measurement data will not get lost. If there is a failure of the data transmission, the measured values can be stored on different hard disks and retransmitted later. The redundant design greatly reduces the risk of failure or data loss.

4 Localization

Reliable localization is one of the most important parts of the robot system. On the one hand, it ensures that the measurement results can be correctly assigned and, on the other hand, MARWIN must be able to find its charging station safely. Repeat accuracy also plays a very important role for meaningful measurement results. The localization is based on the measurement data of the laser scanner and on the odometry data of the wheels. The odometry data are determined from the individual speeds of the wheels and from this the resulting motion of the robot.

In general, a distinction is made between global and local localization. The second has already been researched and discussed many times. In local localization, the robot begins by summing the odometry or laser data to determine its current position from the starting position. It estimates its relative position to the start. For absolute positioning, global localization is added. This topic is far more complex, as the robot must be able to record more information about its environment and thus determine an absolute position. At this point, the "kidnapped robot" problem should also be mentioned [10].

The entire IT hardware of the robot is linux-based and builds on the "Robot Operating System" (ROS)[9]. The main components of the software framework of ROS are hardware abstraction, message exchange, packet management and software libraries. The system is divided into the actual basic system ROS and a selection of additional packages which extend the basic system by individual capabilities. ROS is published under the BSD license and is thus open-source. How exactly ROS works and is implemented has already been described several times, see e.g. [11, 12, 13].

4.1 Environmental conditions

The total length of the XFEL tunnel is about four kilometers. The robot-assisted radiation measurements are carried out for two sections each one kilometer long. As shown in Figure 3, the tunnel is straight. The robot moves along a corridor about 1.4 m wide along the tunnel. The charging stations described in section 3 are located underneath the beamline. In order to prevent damage to the research facility when entering the charging station, precise positioning is crucial. The robot must not collide with the system in any situation. As shown in Figure 3, QR codes are mounted along the tunnel on a rail at 10-meter intervals. The absolute position is encoded in the QR codes.

Furthermore occurs during operation of the accelerator facility ionizing radiation. This consists mainly of gamma and neutron radiation. Since the robot is used

[8] see https://www.hokuyo-aut.jp/search/single.php?serial=167

[9] see http://www.ros.org

during operation, it can happen that one bit flip over due to radiation, software errors can occur.



Figure 3: View along the XFEL tunnel

4.2 Solving approaches

It is important for the robot to be able to determine its absolute position in the tunnel because it must travel to exact coordinates for targeted measurements and loading maneuvers. Therefore, global localization is essential. For this reason, a camera was attached to the robot, so QR codes can be scanned by passing. The position information is encoded in the QR code, the edge length is 5 *cm*, and the average distance between QR code and CCD camera is 120 *cm*. The image data of the QR camera are read in and converted to a ROS format. The image is then decoded by the ROS package *zbar_ros*. The result is a string that contains the tunnel position.

Since the QR codes are 10 meters apart on average, further localization is required. There are basically two approaches to this:

- Localization without known map (SLAM)
- Localization with known map

So-called "Simultaneous Localization And Mapping" (SLAM) algorithms work on the principle of adding up the smallest trajectories and thus drawing conclusions about the relative position. There are different sources of information. In the case of MARWIN, a SLAM method was tested and evaluated with the existing sensors.

Likewise, the localization was tested on the basis of an existing map. However, this method requires an already existing as accurate map. These exist theoretically in the blueprints of the XFEL research facility, but there are many details that actually look different (eg mobile pumps, work platforms, smaller implements, tools, etc). For this reason, a map was generated by means of QR codes and SLAM methods. Based on this map a pure map based localization was tested.

4.3 SLAM results

Different ROS packages were used for the SLAM test. During the software research of possible SLAM algorithms, it was found that no software package allows localization correction by external sensors such as the QR camera. The information of the QR codes can therefore not support the localization in the SLAM approach. The condition of supporting laser and odometry data has fulfilled two software packages, *gmapping* by OpenSlam[10] and *hector_slam* by TU-Darmstadt[11]. Both build on so-called particle filter which has been described in several papers [14, 15, 16]. In the application, they differ in that *hector_slam* also works without odometry data. Furthermore, they differ in the generation of the maps. The software *gmapping* starts with a freely definable initial map size. When the robot reaches the limits of the initial map, it is extended by another unknown block. The software *hector_slam* is different here. The map size must already be known at the beginning, as well as the map must be square. It is not dynamically expanded during runtime. If the robot reaches one of the map boundaries, it loses itself and the localization fails in this area.

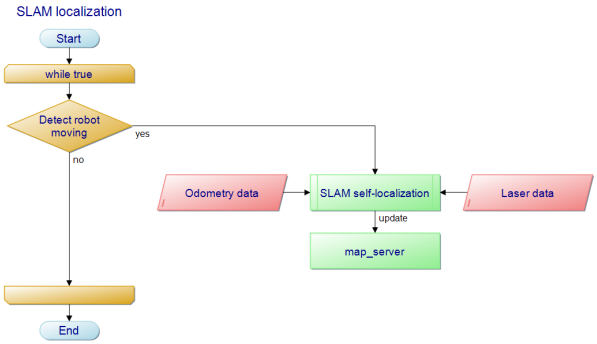


Figure 4: Flow chart of SLAM localization

Figure 4 shows the procedure for the SLAM approach. As the robot moves, this map will be expanded with obstacle information through the SLAM algorithm. The software receives the information from the odometry and laser sensors, weights them and estimates the probable pose.

4.3.1 Using gmapping

First tests with *gmapping* showed that the laser scanners are not arranged with sufficient accuracy (see fig-

[10] see http://wiki.ros.org/gmapping
[11] see http://wiki.ros.org/hector_slam

ure 5). However, this is not a *gmapping* error, but a misalignment of the laser sensors. The resulting map clearly shows a curvature that is not present in reality. The XFEL tunnel is straight.



Figure 5: Incorrectly set laser data lead to curvature of the straight tunnel

Then the test was repeated for a distance of 50 meters and the laser data was recorded. In a simulation a correction factor could be determined over several iterations and thus the arrangement of the laser scanners could be corrected by software. When creating maps with an edge length of 1000 meters and more, a limit was detected. It seems like there is a maximum size of the map, which depends on the chosen resolution. It was found that at a resolution of 20 *cm* per pixel the entire tunnel section (about 1000 *m*) can be mapped. By increasing the resolution to 10 or 5 *cm* per pixel, the dynamic map extension stops after some successful expansion. This has the consequence that *gmapping* is usable in this case only at a maximum resolution of 20 *cm* per pixel. However, the positioning accuracy is worse than expected. The rough map has the consequence that the robot can not approach its measuring positions exactly. Likewise, a precise retraction into the charging station is not possible.

4.3.2 Using hector_slam

As already described, the map size is already defined at program start and is not dynamically expanded during runtime. In addition, the map is necessarily square, since only the edge length of the map can be specified. It has been found that the memory used and the resulting computation cost is significant for a 20000 *px* square map (1000 *m*/0.05 *m* px^{-1}). Therefore the software was adapted and the possibility of a rectangular map implemented. This allowed the map to be reduced to 20000 *px* length and 200 *px* width. It also increased performance and reduced processor load.

Our first test purpose was the reproducibility of the measuring points. Therefore, we let MARWIN perform, for three different distances, 20 times the same test drives autonomously. The first given range was 50 *m*, the second 100 *m* and finally 200 *m*. The tests were always starting from the same point at 280m of the XFEL. The distance travelled in real world was compared against the distance the robot was estimating. Finally, the deviation of the position alongside the tunnel was determined. The results are shown in figures 6. It seems that the localization error is increasing with longer traveled distances.

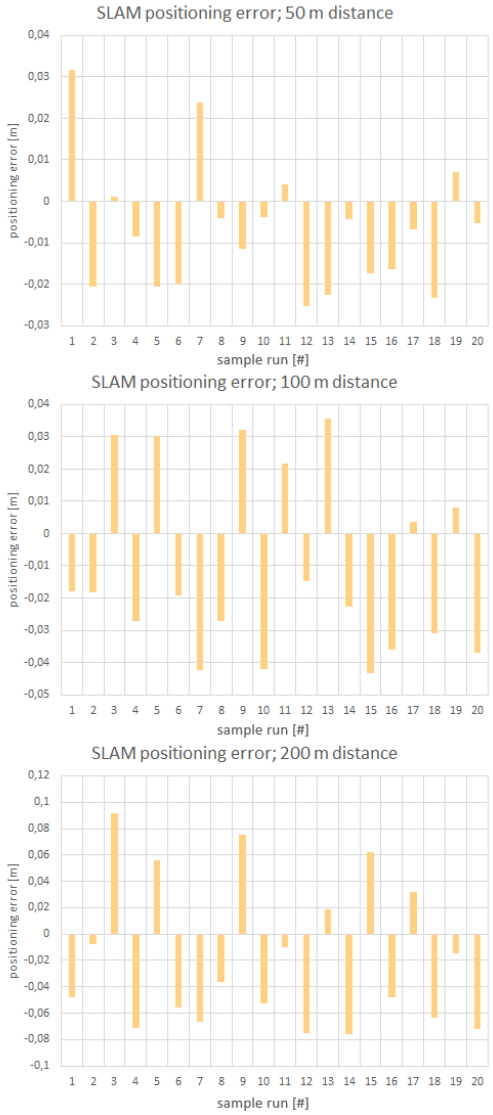


Figure 6: Experimental results of slam test drive (50 to 200 *m*) using hector_slam in the XFEL tunnel

Since the robot has been delivered, the system is constantly being further developed and improved. For a different section in the tunnel (1100 *m* to 2100 *m*) a second robot with the same sensor setup was built. Contrary to expectations, however, the position accuracy has deteriorated considerably here. After a distance of 900 *m* the robot had a positioning error of 9.63 *m*. Some areas in the second tunnel section lead to positioning problems.

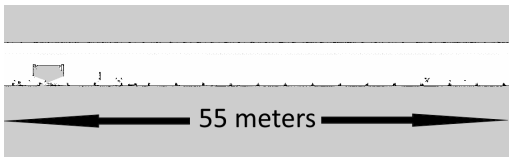


Figure 7: Area with less obstacles in XFEL lead to localization errors

As shown in figure 7 areas were passed with nearly no objects. Within these areas *hector_slam* had problems for determining the travelled distances. For bet-

ter results the covariances of odometry data had to be optimized. The covariances give the odometry data a weighting and thus describe a probability of correctness. The procedure of the optimization can be considered in detail in [17]. As a result the error after travelled distance of 900 *m* could be reduced from 9.63 *m* to 2.43 *m*.

4.4 Static map results

The final map of the SLAM tests described above was taken for static map tests. As described, there is an error of 2.43 *m* after 900 *m* travelled distance, resulting in an error of 0.27 %. The length of the map has been corrected by this value in order to approximate it linearly to reality.

The software for the static map tests is included in ROS as a software package from the navigation stack, called *amcl*[12]. As shown in figure 8, it requires sensor input of 2D-LIDAR data, odometry data and a given static map. Optionally, the position estimation can dynamically be set to a given point in the map. In this way, the QR data are used.

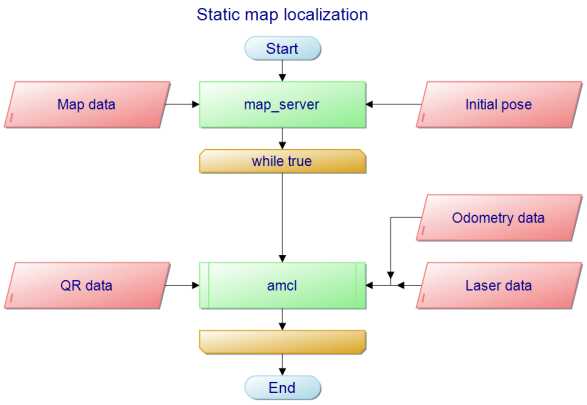


Figure 8: Flow chart of static map localization

Tests for the different distances (50 *m*, 100 *m* and 200 *m*) were repeated as done before in SLAM tests. The results are shown in figure 9. The tolerance in the positioning accuracy is about 5 *cm*.

4.5 Influence of radiation

Due to the redundant structure of the computer units a failure of a unit can be compensated. However, it has been found that the neutron radiation measurably affects the laser sensors. Depending on the intensity of the radiation, it happens that individual measuring points of the laser scanner are falsified. This expresses itself concretely in the fact that the distance in this point is zero.

Due to the fluctuating radiation conditions in the XFEL, it is difficult to carry out targeted radiation tests and to condition the robot system with defined radiation doses. Nevertheless, initial findings were obtained. For this purpose, MARWIN measured weekly the radiation doses of the accelerator system and the number of falsified measuring points. These data were related to each other (see figure 10). The radiation was recorded at a distance of about 80 *cm* from the laser scanners.

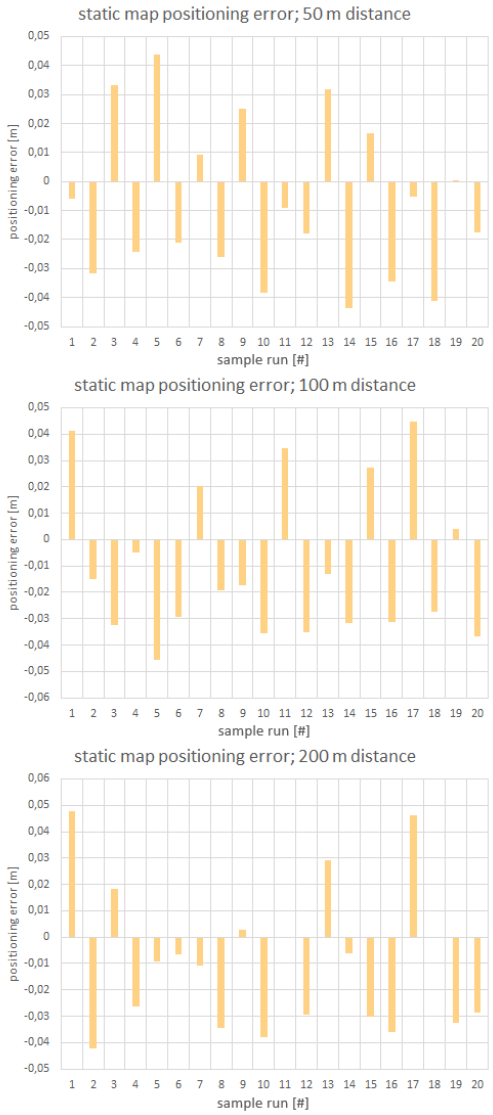


Figure 9: Experimental results of static map test drive (50 to 200 *m*) in the XFEL tunnel

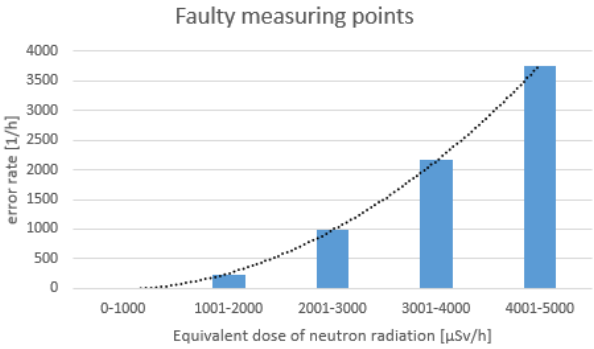


Figure 10: Faulty measurement points for selected groups of neutron radiation doses

[12] see http://wiki.ros.org/amcl

5 Conclusions

The SLAM tests have shown that the positioning error increases with the distance traveled. The SLAM approach with *hector_slam* has great results in map sizes up to 200 m. After that, the accumulated errors become larger, so that the localization error becomes too large for the entire tunnel distance of 1000 m and is therefore not sufficient for the requirements of MARWIN. An unconfigurable relationship between map size and resolution in *gmapping* also leads to insufficient positioning accuracy for MARWIN. In smaller environments, such as the office, these mistakes are not really crucial. However, the accumulation of small localization errors, especially in larger and wide environments, has considerable influence on the positioning accuracy.

The second approach to localization with a static map is a way to move robots in large areas with high positioning accuracy. Due to the positioning tolerance of a maximum of 5 cm, which is related to the resolution of the map, the absolute error is minimal even 1000 m distances traveled. In comparison, the error in the SLAM approach after this distance was 2.43 m.

The combination of creating a map by SLAM method and its linear correction for tunnel-like structures seems to be a good basis for localization methods using static maps.

However, the use of static maps also has the disadvantage that the robot system is no longer able to respond to dynamic changes in the environment. If too many details change in the environment, this can lead to the total loss of localization. This event was detected with tools and mobile equipment left in the XFEL research facility. SLAM approaches are more robust in this scenario.

With the described static map approach the robot system MARWIN is successfully in use. Several times during the operation of the accelerator facility, inspections and autonomous radiation measurements were carried out. For a long-term view of the radiation conditions, a weekly snapshot of the entire research facility is created autonomously by MARWIN.

The influence of neutron radiation on the measurement data of the laser scanner has no noticeable effect on the positioning accuracy. This may be because the number of faulty measurement points per time interval is a fraction of the total number of measurement points. According to the data sheet of the laser scanner used, about 155 million measuring points are recorded per hour, whereas in the dose range of $4001 - 5000\ \mu Sv\ h^{-1}$ only about 3700 incorrect measuring points occur. However, should the radiation dose continue to increase, there could be a significant deterioration on the accuracy of localization due to the tendency of the faulty measurement points.

5.1 Further work

The robot system is constantly being improved and developed further. In the specific case of localization, further approaches could be considered. Currently relatively new but very promising is a new approach called *Cartographer*. This software package is developed by Google and is also based on the SLAM method. Larger environments should be able to be mapped with it [18].

In the long term, further cooperations are to be concluded in order to promote the expansion of other research facilities. It also has a great focus on the development of a manipulator in order to make minor repairs and interventions in the operation of the accelerator system.

References

[1] A. Dehne, N. Möller, and T. Hermes, "Marwin: a mobile autonomous robot for maintenance and inspection in a 4d environment," in *2017 International Conference on Research and Education in Mechatronics (REM)*, Sept 2017, pp. 1–5.

[2] J. H. Zhou, J. Q. Zhou, Y. S. Zheng, and B. Kong, "Research on Path Planning Algorithm of Intelligent Mowing Robot Used in Large Airport Lawn," in *2016 International Conference on Information System and Artificial Intelligence (ISAI)*, June 2016, pp. 375–379.

[3] Y. Zhou, R. Sun, S. Yu, J. Yang, and L. Sun, "An obstacle avoidance method based on non-radial arrangement of distance sensors for vacuum cleaning robot," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2016, pp. 608–612.

[4] D. Portugal, L. Santos, P. Alvito, J. Dias, G. Samaras, and E. Christodoulou, "SocialRobot: An interactive mobile robot for elderly home care," in *2015 IEEE/SICE International Symposium on System Integration (SII)*, Dec 2015, pp. 811–816.

[5] F. He, Z. Du, X. Liu, and Y. Sun, "Indoor dangerous gas environment detected by mobile robot," in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2009, pp. 396–401.

[6] C.-T. Chiang, "A versatile gas/vision tracking robot for security system applications," in *2012 IEEE Conference on Control, Systems Industrial Informatics*, Sept 2012, pp. 50–53.

[7] I. Kruijff-Korbayov, L. Freda, M. Gianni, V. Ntouskos, V. Hlavac, V. Kubelka, E. Zimmermann, E. Zimmermann, H. Surmann, Dulic, Rottner, and E. Gissi, "Ground and aerial robots in earthquake-response in amatrice, italy: a field report," in *Proceedings of the 2016 IEEE International Symposium on Safety, Security and Rescue Robotics (SSSR)*, Lausanne, Switzerland, October 2016.

[8] K. Kershaw, F. Chapron, A. Coin, F. Delsaux, T. Feniet, J. L. Grenard, and R. Valbuena, "Remote inspection, measurement and handling for lhc," in *2007 IEEE Particle Accelerator Conference (PAC)*, June 2007, pp. 332–334.

[9] B. I. Erland, "Wheels for a course stable selfpropelling vehicle movable in any desired direction on the ground or some other base," Patent US3 876 255, 1975.

[10] I. Bukhori, Z. H. Ismail, and T. Namerikawa, "Detection strategy for kidnapped robot problem in landmark-based map monte carlo localization," in *2015 IEEE International Symposium on Robotics and Intelligent Sensors (IRIS)*, Oct 2015, pp. 75–80.

[11] M. Köseoğlu, O. M. Çelik, and Ö. Pektaş, "Design of an autonomous mobile robot based on ros," in *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, Sept 2017, pp. 1–5.

[12] Q. Xu, J. Zhao, C. Zhang, and F. He, "Design and implementation of an ros based autonomous navigation system," in *2015 IEEE International Conference on Mechatronics and Automation (ICMA)*, Aug 2015, pp. 2220–2225.

[13] J. Boren and S. Cousins, "Exponential growth of ros [ros topics]," *IEEE Robotics Automation Magazine*, vol. 18, no. 1, pp. 19–20, March 2011.

[14] G. Grisetti, C. Stachniss, and W. Burgard, "Improved techniques for grid mapping with rao-blackwellized particle filters," *IEEE Transactions on Robotics*, vol. 23, no. 1, pp. 34–46, Feb 2007.

[15] G. Grisetti, C. Stachniss, and W. Burgard, "Improving grid-based slam with rao-blackwellized particle filters by adaptive proposals and selective resampling," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, April 2005, pp. 2432–2437.

[16] S. Kohlbrecher, O. von Stryk, J. Meyer, and U. Klingauf, "A flexible and scalable slam system with full 3d motion estimation," in *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*, Nov 2011, pp. 155–160.

[17] P. Eberhard and Q. Tang, "Sensor Data Fusion for the Localization and Position Control of One Kind of Omnidirectional Mobile Robots," in *Multibody System Dynamics, Robotics and Control*, pp. 45–73, Springer, Vienna, 2013.

[18] W. Hess, D. Kohler, H. Rapp, and D. Andor, "Real-time loop closure in 2d lidar slam," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1271–1278.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 354-364 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Modified HOG Descriptor-Based Banknote Recognition System

Tamarafinide Victory Dittimi*, Ching Yee Suen

Center for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, Canada

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 11 October, 2018
Online: 18 October, 2018

Keywords:
Banknote recognition
Histogram of gradient
Support vector machine

ABSTRACT

This research presents a banknote recognition and counterfeit detection system that proposes a pre-processing approach before creating a new feature set by extracting Speeded-Up Robust Feature, creating image patches from its vertices and then computing the Histogram of Gradient descriptors from cells within the patch boundaries. A feature reduction approach is also designed based on a high correlation and Low variance filter to reduce the feature set by calculating the intra-feature value variability to ascertain the similarity between the elements to prune out highly correlated values and variances using a fixed threshold. For experimental evaluation, additional analysis was conducted using Principal Component Analysis, Bag of words, Low Variance filter, and High Correlation Filter method. Furthermore, the classification was performed using 10-fold cross-validation based One-Versus-One Support Vector Machine. Additionally, the developed technique was tested on counterfeit banknotes to authenticate genuine and fake currencies. Lastly, point matching method was applied to detect our suggested feature vectors to match images containing multiple notes to distinguish and determine the value of the bills.

1. Introduction

Technological advancement has made life more comfortable, but it also has its downside as criminals have taken advantage of this progression. Initially, only printing houses had access to the technology needed to produce banknotes but today with the aid of a personal computer and a good laser printer people can duplicate paper money easily in their residence. Banknote counterfeiting is one of the core crimes on earth. The Apex bank of every country is responsible for providing the security features and in Nigeria it is the responsibility of the Central Bank of Nigeria (CBN). Every Nigerian banknote contains several security features that can be used to identify and recognize different note denominations [1].

Over the years, differentiating between fake and real currency banknotes has become quite tricky due to the fast-scientific progress in color printing, cloning, and imaging and because fake notes are now reproduced via top-notch technology that uses security paper [2]. In the same view, counterfeit banknote detection has become an essential task in billing machines. Its implemented by using image analysis methods (transmittance analysis, reflectance analysis) with different light spectra (visible, infrared and ultraviolet). The drawback of this system is its high cost and its usage with only Automated Teller Machines (ATM), where an elevated level of reliability is required. Anti-counterfeiting banknotes security features can be classified into three categories: printing, ink and substrate features. Printing features like intaglio printing, micro letters, serial number, signatures, Guilloches, Ink features such as UV glowing ink, pearl ink, infrared ink, color-shifting inks, and substrate features like planchets, security thread, see-through registers, and watermark [3].

This research presents a paper money recognition system that includes multiple banknote identifications and Counterfeit note detection that could aid in reducing banknote counterfeiting offenses and positively influence a country's economy. The rest of the article is divided into the following sections: Section 2 discusses the research background and surveyed literature. In Section 3, the proposed system including its architecture, design, and stages of the approach is studied. Section 4 debates the analysis and discussion of the results of the system functionality, and the conclusion of the research and future work is given in Section 5.

2. Background & Literature Survey

Banknotes are utilized in every business venture in as much as electronic currencies is becoming identified as a substitute legal tender, the note is still dominant and central in the day to day dealings. Currency Recognition and detection is a critical problem in computer vision. Although, electronic and online currencies exist banknote is still the preferred option in daily transactions such

*Tamarafinide Dittimi, Concordia University, t_dittim@encs.concordia.ca

as vending machine, the banking system, banks, shopping mall, railway ticket counter, ATM, vending machines, and currency exchange service. Furthermore, there are cases of money bundles withdrawn from ATMs contained a few counterfeit bills although banks are unwilling to accept these omissions. Even though this problem is majorly unavoidable, they are not also able to provide all branches with the needed counterfeit detection system [4].

One of the most straightforward methods that have been applied in paper money recognition system is the use of visible features found in notes like the size, texture, and color of banknotes. All banknotes have pictures and patterns impressed on them, such as portraits of historical persons or images of landmark buildings which are characterized as textures extracted as features using texture-modeling techniques. The drawback of this approach is that banknotes are mistreated so could get damaged and torn due to regular use leading to alteration of the notes thereby hurting the recognition rate [5]. On the other hand, [6] extracted the geometric features (length and width) and employed it in classifying bills with varying sizes such as Italy and Chinese banknotes. Although, this approach was simple and fast; it had significant limitations including its effectiveness for slant, stain, and tear. Furthermore, it was unsuitable for some banknote such as the US Dollars, which had several denominations of the same size.

Furthermore, many countries employ colors for differentiating banknotes although they could wear or tear it still retains its color [7]. [8] designed a system that identified the denomination of the Indian currency note based on its color. Likewise, another author also developed a technique that converted the image from RGB color space to HSV color space and then extracted the color feature. Their justification for employing the HSV color space was because of its nearness to the human conceptual understanding of color [9]. Additionally, a researcher recommended an approach based on characteristic point extraction of a unique part of the Euro paper currency that possesses the same color has also been conducted [10]. Similarly, [11] devised a US banknote counterfeit detection system that extracts its color feature and classifies the currency using fuzzy Hamming distance. Despite its high performance, if the system were tested on Euro banknotes. It would not be able to detect forgeries since its detection is based only on the color feature. Most countries have different colors distributed over the surface of its banknotes including the Nigerian Naira; all have only one dominant color. For instance, the dominant color of 10 Naira note is “Red.” Although, some banknotes have the same dominant color due to wear and tear the color quality may degrade; hence leading to difficulties in using the color feature for automatic recognition. Also, using only the color as a discriminative feature may fail because some banknote denomination does have the same dominant color [12].

Likewise, [13] proposed a banknote classification technique based on discrete real-valued wavelet transform (DWT), they calculated the feature vector using the statistical properties of decomposed coefficients from the approximation and detail sub-bands. The technique was applied successfully in spatial and frequency domain, although it does not shift invariant and has no phase information. Another researcher also devised a new method called a hyper-complex wavelet transform (CWT) that is shift invariant and has a single phase. However, it also had its limitation like its ambiguity when used for 2-D signals [14]. Furthermore, an invariant shift extension of DWT and CWT called the quaternion wavelet transform (QWT) was also adopted. The technique had three phases and one magnitude, the first two stages were on the local image shifts, and the third focuses on image texture information [15]. [16] also suggested a new feature extraction system using the quaternion wavelet transform (QWT) that combined the partial and total Hilbert Transform (HT) using quaternion algebra to form one magnitude and three phases. It captures the statistical characteristics of QWT coefficients using the generalized Gaussian density (GGD) and classifies the banknote using the neural network. Furthermore, [17] presented a banknote recognition system that used a wavelet transform for extracting features and classified the features using Multi-layer perceptron (MLP) neural network; the method experimented on Persian paper currency. In another research work, a feature extraction technique was developed based on the Discrete Wavelet Transform (DWT). They employed Daubechies 1 (DB1), and the approximate coefficient matrix of the transformed image is extracted from the estimated efficient model and used for recognition, classification, and retrieval of currency notes [18].

Additionally, [19] formulated an approach that uses the edge and corners of an object in detection, the system was invariant to image rotation but was not able to handle issues associated with image scale. [20] then presented another technique that could automatically determine the scaling of the objects, it computed the Hessian matrix and used the Laplacian operator to locate the block like feature points in the image. Similarly, [21] also conducted further research on automatic scaling selection and implemented a Harris-Laplace detector in addition to the Hessian-Laplace to improve the automatic scale selection process. [22] also developed a SIFT method that finds the limits of the scale space in an object. The technique was invariant to scale and rotation and was very useful in extracting interest point descriptors which can be utilized in matching and classifying objects. [23] devised a method for the classification of Slovak banknotes based on SIFT detector and descriptors. Although, the approach posed its issues which were its computational complexity leading to high processing time limiting its acceptability in real-time applications.

In the same view, [24] also developed a colored SIFT approach that combines the color feature and the geometric data using the SIFT algorithm. The Mikolajczyk and Schmid framework was further applied to normalize introduced a nonlinear SVM method that juxtaposed images of different color space in addition to combining color space descriptors to evaluate the accuracy of the approach [25]. On the other hand, speeded up robust features (SURF) was created, the process was designed to improve on the SIFT method, the procedure retains the robustness of SIFT but with a decreased descriptor size and shorter processing rate and time.

SURF is also an interest point detector that is robust to scaling and rotation. It creates a unique, repetitive and vigorous process for capturing the interest points. The technique creates an integral image and uses it in place of the original image to diminish the processing time. It then uses the Hessian matrix to detect the highest and lowest second derivative value after this is achieved the interest point orientation is detected and rotated to align with the orientation of the predefined direction to complete the SURF descriptor creation [26]. A researcher applied SURF descriptor in

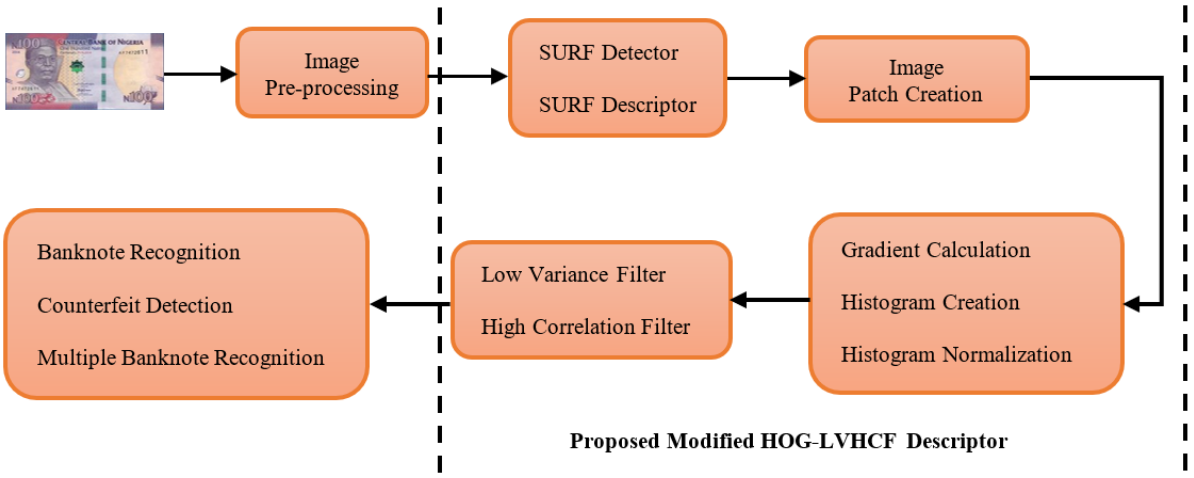


Figure: 1. Proposed Modified Hog-LVHCF based Banknote Recognition System

banknote recognition; the system was used to match reference regions with query images. It presented promising results and was robust to handle partial distortions as well as worn or wrinkled bills [27].

3. Proposed Methodology

In this research, we implemented a procedure for enhancing and cleaning up banknote images. It involves converting to grayscale then employing Gamma Correction and Histogram equalization to improve the image luminance and contrast of the intensity image. Furthermore, Weiner filter was applied to reduce the amount of noise after that the boundaries of the banknote are computed using Blob detection, cropped out and Gaussian pyramid utilized in standardizing the note. We also created a new feature set by extracting the Histogram of Gradient (HOG) Descriptors from image patches generated from the vertices of SURF interest points. After the HOG features have been obtained, the vector size may differ for each image in addition to being computationally demanding during processing. To minimize its effect on recognition/detection rate while maximizing the reduction of Feature dimensionality. A combination of dimensionality reduction techniques based on the descriptor correlation and variance is applied to reduce the Vector scale. A descriptor filter analysis using Pearson's Product Moment Coefficient ρ and variance σ^2 is employed to determine the best set of features for classification.

The approach answers two fundamental questions, determining the relevance of the descriptor to the feature set and its redundancy regarding other associated descriptors. Finally, we applied our technique to several datasets using OVO-SVM and Putative point matching to investigate the efficiency of the descriptor showing the effectiveness of Modified HOG/LVHCF descriptor for banknote identification, counterfeit detection, and multiple notes recognition. Figure 1 shows the phase of the proposed Modified Hog-LVHCF based paper money recognition system.

3.1. Image Pre-processing

During image acquisition, the irrelevant background image is also captured. This unwanted information must be removed to identify and recognize the notes. Furthermore, digitized banknote images sometimes have discontinuous particles, noise, illumination, positioning, and unanticipated boundaries. Also, when matching the image, if the banknotes are of varied sizes it could affect the recognition rate. In this research, image pre-processing plays a fundamental step as it involves enhancing and cleaning up the images for efficient feature extraction. Thus, in this research, a new combination of techniques for image pre-processing is presented, and a flowchart of the procedure is shown in Figure 2.

Step 1: The image was first converted to grayscale by calculating the weighted sum of the three channels of the RGB image and representing the result as a matrix whose fragments are assigned values based on how dark or bright the color of the pixel is at the corresponding position. The value allocated to the pixel during grayscale conversion ranges from 0 to 255 representing the brightness of the image.

Step 2: Gamma Correction was applied to improve the image luminance by enhancing the quality of the image in dark or blurry areas while also compressing the intensity of bright regions.

Step 3: Histogram equalization was then employed in enhancing the contrast of the intensity image, the approach redistributes the image histogram to ensure that the brightness value is constant across the image.

Step 4: Weiner filter was applied to reduce the amount of noise present in a signal comparing with an estimate of the desired noiseless signal.

Step 5: Background subtraction was also achieved by first finding all blobs in the image using Blob detection, select the boundary coordinates of the blob containing the banknote and cropping out the note using the detected blob edges.

Step 6: Banknote denominations sometimes are of varied sizes, and the note can also change due to the acquisition manner thus image reduction and standardization were achieved using the Gaussian pyramid.

3.2. Feature Extraction (Proposed Modified HOG Descriptor)

Feature extraction is a crucial step in image processing, the selection of features is vital in determining the recognition of an object, and an ideal feature set should be stable, invariant to rotation, scale, and illumination changes. Several feature

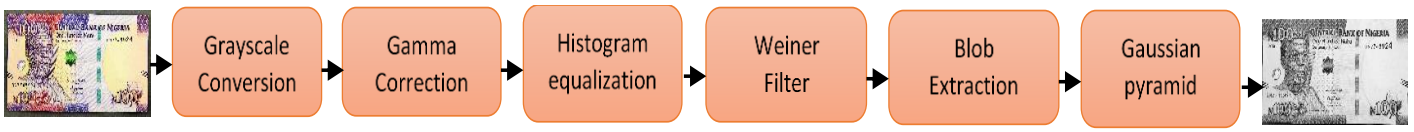


Figure 2: Block diagram of the procedures undergone in the Image Pre-processing

extraction methods have been proposed over the years. In this research, several useful descriptors and feature reduction approaches were investigated and compared to the proposed Modified HOG descriptor and Low Variance High Correlation Filter. In this study, six different descriptors were employed: Scale Invariant Feature Transform, speeded up robust features, gradient location and orientation histogram, Histogram of Gradient, shape context, and the proposed Modified HOG Descriptor. While to reduce the processing time several feature reduction methods were also implemented: Principal Component Analysis, Bag of Words, High Correlation filter, Low Variance Filter, and compared to the developed Low Variance and High Correlation Filter Algorithm. The proposed Modified HOG based descriptor technique is split into four phases: SURF point detection, SURF descriptor extraction, image patches creation from SURF points, and HOG feature computation.

3.2.1. SURF Point Detection

The SURF interest point detection is separated into four phases: integral image generation, calculation of the matrix, construction of the scale space and localization of the feature point.

Integral Image Generation: The Integral image generation was introduced to reduce the computational complexity of the square like convolution filters used in the future, it is mapped from the original image. The integral image is then computed by the sum of all pixels in the input image I within a given rectangle sector which can be deduced by Equation 1.

$$I_{\Sigma}(x,y) = \sum_{i=0}^{x}\sum_{j=0}^{y} I(i,j) \qquad (1)$$

After the integral image has been calculated, three additions are carried out to compute the summation of the intensities over any four-sided vertical area consequently the computational time is not dependent on the image size using Equation 2.

$$\sum = I_{\Sigma}(A) + I_{\Sigma}(D) - \left(I_{\Sigma}(C) + I_{\Sigma}(B)\right) \qquad (2)$$

Computation of the Hessian Matrix: The box filters are used as an approximation of the second-order Gaussian derivatives, this is computed on the integral image calculated earlier rather than the actual image hence reducing the processing time. A 9x9 square-like filter was employed in the computation of the Gaussian denoted by with a scale of 1.2 which led to the creation of a blob response map over different scales in the image. A set of simple weights w were applied to the square-like sector to reduce the computational overhead.

Construction of the scale space: The SURF approach is scale invariance, it applies the same filter at different sizes to the image at the same speed in parallel. The length of the dark lobe can only be increased by an even number of pixels to guarantee the presence of a central pixel (top). The scale space is categorized as a series of octaves; an octave is a set of filter response maps gotten by convolving each input image with a filter in different sizes. For an octave layer with a size N * N, the scale σ, the convolution of each layer in the image is computed on the response map using equation 3.

$$\det(H_{approx}) = D_{xx}D_{yy} - (wD_{xy})^2 \qquad (3)$$

Localization of the interest point: The final step in the interest point detection is localizing the image over the different scales, a threshold is set for removing weak responses in the map. Several thresholds are checked until a middle ground is reached that takes into consideration the number of interest points and the strength of the detected interest point. Furthermore, a 3-dimensional scale space is applied to the non-maximum suppression, and only points bigger than all the neighbors is selected as the given interest points. Finally, scale space interpolation is applied to obtain the best interest points [28].

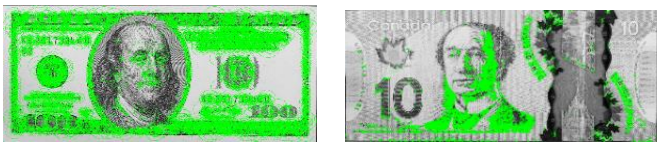
3.2.2. SURF Descriptor Extraction

SURF Descriptor computation is divided into three phases: Assignment of the interest point orientation, Computation of the descriptor using Haar Wavelet transform and Feature matching using Fast Indexing.

Assignment of the interest point orientation: Assignment of Interest points orientation is needed to attain rotation invariance, so Haar Wavelet transform is used at each interest point (x, y) to determine the direction of a circular neighborhood in addition to its scale. After computing the wavelet transform the feature point is weighted using a Gaussian filter, the dominant orientation is decided by calculating the summation of all the horizontal and vertical responses within the sliding windows; the most extended vector is then selected as the local orientation vector for the feature point.

Computation of the descriptor using Haar Wavelet transforms: The previous step used a box area centered around the point of interest in determining the orientation. In this phase, each sector is further divided into a 4x4 square like sector, to improve the reduced the geometric distortions, and localization inaccuracies that could occur the interest pointed are first weighted with a Gaussian filter then the Haar wavelet transform is calculated horizontally and vertically for each 5x5 paced interest points. Furthermore, the computed wavelet transform is then summed for each sector and use as the first level entry in forming the feature vector set. Moreover, to capture the change in intensity after each interest point the absolute value of the responses is also computed to create a 4-dimensional descriptor vector for each sub-sector. Finally, the sectors which are made up of a 4x4 sub-sector is concatenated resulting in a descriptor with a length of 64 dimensions.

Feature matching using Fast Indexing: This phase presents an additional functionality to the matching stage, it uses a fast indexing technique based on the previous computed Laplacian in the Haar wavelet transform stage. The Laplacian value can be used to differentiate lighter blobs on darker areas from darker blobs in brighter areas. Hence, used in removing features that do not share the same type of contrast thus reducing unstable interest points from the feature vector set thus reducing the processing time and as the Laplacian has already been calculated earlier it to does not add any extra computation overhead. Figure 3 presents SURF point descriptors of sample USD and CAD paper money.



a) b)

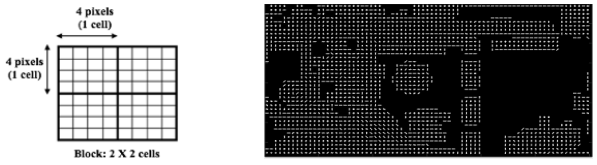
Figure 3. Detected SURF descriptors of a) the US, and b) Canadian Dollar

3.2.3 *Creation of Image Patches*

An image patch is designed from polygons defined by the coordinates of its vertices; it consists of one or extra polygons that might or might not be connected. In this research, the 43 x 43 patch is formed by specifying the coordinates of each unique vertex of the SURF square-like descriptors and a matrix that defines how to connect these vertices to create the faces. This technique is preferred for SURF points as its patches multifaceted, so vertices shared by more than one face need be defined only once thereby requiring fewer data to design the patches.

3.2.4 *HOG Descriptor Computation*

Histogram of Gradient (HOG) feature is a viable feature extraction technique with the ability to select feature points irrespective of the banknote image view. It is based on the histogram of pixel gradients neighbors for image sections or blocks thus it is hardly affected by changes in lighting conditions or image geometry. The HOG feature extraction technique can be grouped into three main stages: Calculation of Gradient for each pixel, computing its Histogram and by combining the histogram of all cells in a block to normalizing the histogram [29].



a) b)

Figure 4. (a) Cell and block module (b) Sample HOG on Naira note

The feature depicts the gradient dispersion on the different positioning of a banknote image and is conducted to obtain the dimension and texture feature of the banknote. The HOG algorithm partitions all note image into small evenly-dimensioned sections called cells comprise several pixels which individually have a weight computed by getting the gradient magnitude and histogram. The histogram shows the dispersion of the pixel taking into cognizance the orientation of the gradient. Cell histogram is offered as a vector that is merged to get the HOG feature of the overall banknote image [30]. Figure 4 shows the terminology of the HOG feature extraction. Finally, Table 1 shows the HOG method specification that provided the best result in this research. The Cell size of 16, four cells per block, 1 to 1 block overlap, nine bins, and 18x1 bin centers. The number of HOG feature is different for each Dataset because it is chosen based on the feature size that presented the best recognition and detection rate.

Table 1. Specification of the HOG Descriptor

Specification	Naira	USD	CAD	Euro
SURF Descriptor	124864	196240	209002	205855
Cell Size	16	16	16	16
Number of cells per block	4	4	4	4
Block Overlap	1,1	1,1	1,1	1,1
Number of bins	9	9	9	9
Number of HOG extracted	38102	56109	76464	70136
Bin centers	18 x 1	18 x 1	18 x 1	18 x 1

3.3. *Feature Reduction (Proposed LVHCF)*

After the HOG features have been extracted, the vector size is computationally demanding. Therefore, to unify and reduce the feature dimension; the techniques aim to minimize its effect on recognition/detection rate while maximizing the reduction of Feature dimensionality. The experiment conducts a descriptor filter analysis using Pearson's Product Moment Coefficient ρ and variance σ^2 to determine the best set of descriptors for classification. The technique answers two fundamental questions: determining which descriptor is relevant or not to the feature set and if the descriptor is redundant or not concerning other related descriptors.

$$\rho = \frac{m(\sum ij)-(\sum i)(\sum j)}{\sqrt{[m\sum i^2-(\sum i)^2][m\sum j-(\sum j)^2]}} \quad (4)$$

$$\sigma^2 = \frac{\sum(x-u^2)}{m} \text{ where } \mu \text{ is the mean} \quad (5)$$

The first question can be answered by just defining a threshold ρ value, but for the second question it is more complicated as it involves analysis of pairwise correlations between all descriptors. To answer both questions before filtering the descriptors we put forward an approach called a Low Variance and High Correlation Filter (LVHCF). In this experiment, it is assumed that if two descriptors are found to be redundant to each other one of them needs to be removed and removing the one that is less relevant to the descriptor-set concept keeps more information to predict the class while reducing redundancy in the data.

Table 2. Feature Reduction on Datasets

Technique	Naira	USD	CAD	Euro
PCA	680	740	690	755
BOW	600	650	650	670
LVF	1905	2805	2293	3507
HCF	1143	1685	2294	2104
LVHCF	581	661	764	701

As presented in Figure 5, given a dataset with M feature set and a descriptor set ds, the algorithm chooses a list of predominant descriptors D_{best} for a given ds. It comprises two phases, the first phase (line 2-17) computes the Pearson correlation moment coefficient p for all pairs of descriptors in the ds using equation 1, extracts relevant descriptors into $S1^i_{list}$ based on the predefined threshold t1. The $S1^i_{list}$ is further processed to remove additional correlated pairs and only keeps predominantly uncorrelated D*q* and *Dp* among all the earlier selected descriptors in the $S1^i_{list}$ (line 13). After one round of filtering all D*q* pairs based on D*p*, the algorithm will take the currently remaining feature right next to *Dp* as the new reference (line 18) to repeat the filtering process. The algorithm stops until there is no more feature to be removed from $S1^i_{list}$. In the second phase (line 19 – 38), the variance σ^2 of each descriptor in the $S1^i_{list}$ is gotten using equation 2. Furthermore, the descriptors are extracted into a new list $S2^i_{list}$ based on the predefined threshold t2 and ordered in ascending order according to their σ^2 values. The list is further processed to remove redundant descriptors and only keeps predominant ones among all the selected relevant descriptors. A descriptor D*p* that has already been determined to be dominant can always be used to filter out other features that are ranked higher than D*p* and have D*p* as one of its redundant σ^2 peers. The iteration starts from the first descriptor in $S2^i_{list}$ (line 26) and continues as follows. For all the remaining descriptors, if D*p* happens to be a redundant peer to a feature *Dq*, *Dq* will be removed from $S2^i_{list}$. After one round of filtering descriptors based on D*p*, the algorithm will take the currently remaining descriptors right next to *Dp* as the new reference (line 37) to repeat the filtering process. The algorithm stops until there is no more feature to be removed from $S2^i_{list}$. The descriptors matching the variance remaining in $S2^i_{list}$ is then used as the final dataset B_{est}. Table 2 presents the feature reduction on the four datasets.

Input: S (D_1, D_2, …, D_m, ds)// training dataset

t1 = 0.287 // a predefined threshold (ρ)

t2 = 0.045 // a predefined threshold (σ^2)

output: S_{best} // an optimal subset

```
1   Begin
2     For i = 1 to M
3       For j = 1 to M
4     Do begin
5       Compute ρ(i, ds) for all Di and Dj
6       if (ρ(i, ds) >= t1)
7         Append Di to S1^i_list;
8     End;
9     Dp = getFirstDescriptor (S1^i_list);
10    Dq = getNextDescriptor (S1^i_list);
11      Compute ρ for all Dp and Dq
12      If (ρ >= t1)
13          remove Dq from S1^i_list;
14              Dq = getNextDescriptor (S1^i_list);
15      Else Dq = getNextDescriptor (S1^i_list);
16      End until (Dq == NULL);
17      Dp = getNextDescriptor (S1^i_list, Dp);
18    End until (Dp == NULL);
19    For i = 1 to M
20    Do begin
21      Compute σ^2(i, ds) for S1^i_list;
22      if (σ^2(i, ds) <= t2)
23        append Di to S2^i_list;
24    End;
25    Order S2^i_list in ascending σ^2(i, ds) value;
26    Dp = getFirstDescriptor (S2^i_list)
27    Do begin
28      Dq = getNextDescriptor (S2^i_list, Dp);
29      if Dq <> NULL)
30        Do begin
31          D'q = Dq;
32            if (σ^2(p, q) <= σ^2(q, ds)
33              remove Dq from S2^i_list;
34              Dq = getNextDescriptor (S2^i_list, D'q);
35            Else Dq = getNextDescriptor (S2^i_list,
Dq);
36          End until (Dp == NULL);
37        Dp = getNextDescriptor (S2^i_list, Dp);
38      End until (Dp == NULL);
39  Sbest = S2^i_list;
40  End;
```

Figure 5. Low Variance and High Correlation Filter (LVHCF) Algorithm

3.4. Classification

The classification stage is the decision segment of a recognition system, and it uses the features extracted in the previous phase. It involves identifying each banknote and assigning it the correct denomination class. Furthermore, the principal function of a pattern recognition system is to yield decisions concerning the class membership of the models with which it is confronted. Support Vector Machine (SVM) is a classification approach used for separating hyperplanes, it was designed in the 1990's for building nonlinear separating function, and real-valued estimation [31].

The Multi-Class SVM-based system uses a binary classification approach for the one versus one technique, it designs a classifier for each class and creates a hyperplane between the classes. Furthermore, the system assigns a class to a test sample based on a distance margin for the classes. Each data set is split into two groups, the training, and test groups. The system used a randomized approach for data selection, and the class of the test data was predicted based on the established class of the training mechanism. The Multi-Class SVM has two application method approaches; One Versus All (OVA) method and One Versus One (OVO) method. OVA differentiate one class from the remaining classes, and the OVO distinguishes each class from one class at a time while ignoring the remaining class [32].

For this research, the system used the OVO technique in combination with error correcting output codes. The error correcting output code divides a Multi-Class scheme into a double class system that constructs a classifier for each binary division. OVO technique compares all classes against each other and builds a binary classifier to distinguish each pair of class while ignoring the remaining classes. When classifying an unknown image to a class, the classifier conducts a vote, and the class with the highest majority is selected as the best choice to assign the banknote [33]. Further analysis was conducted using K-Nearest Neighbor (KNN), Weighted K-Nearest Neighbor (WKNN), Random Forest (RF) Decision Tree, and multilayer perceptron (MLP) Neural network. In conclusion, it is hard to decide which classification method is the best as research has only been able to prove that a

classifier could have a higher recognition and detection rate than others based on a specific problem.

4. Result & Discussion

This system was implemented in MATLAB as its toolbox provides built-in support for SIFT, SURF, HOG, GLOH in addition to the several classifiers. The hardware utilized comprised of a Core i3, 2.4 GHz processor with 4 GB Ram, the operating system used was Windows 8.0 64Bit and image processing techniques, and other related MATLAB toolboxes employed in the development of the system. This research utilized 10-fold Cross-Validation for the analysis of the data. It is a statistical technique used in evaluating predictive models by dividing the data into training and validation sets. The approach divided the dataset into K-fold equal data sizes and crossed over each K-dataset partition in successive rounds such that each data point has a chance of being in the validation process. The Centre for Pattern Recognition and Machine Intelligence (CENPARMI) has developed a database for Banknote Identification (Experiment 1), Counterfeit Detection (Experiment 2), and Multiple Banknote recognition (Experiment 3). The database comprised of eight classes for the Naira Notes, US Dollars had six classes, five classes for the Canadian Dollars, and seven classes for the Euro Bills for both genuine and counterfeit bills [34, 35].



a) b) c)

Figure 6. a) Full notes b) Folded notes, and c) Occluded notes

4.1. Banknote Recognition (Experiment One)

The obverse and reverse of the banknotes were captured using a Canon T2i digital camera. Banknotes suffering from scratches, wrinkle, stains, wear and tear which represent the image quality of the paper monies in the real-world were captured for this project. The statistical distribution of the Dataset was Nigerian Naira, US Dollar, Canadian Dollar, and Euro data sets. It comprised 1790, 1120, 1000, and 1400 genuine banknotes and 1320, 823, 733, and 1020 fake banknotes respectively. Each dataset included images with multiple notes, clutters, overlapping and partial paper money. Furthermore, the datasets were made up of images captured in four directions for each banknote: the front (upward and downward) and back sides (upward and downward).

Figure 6 depicts pre-processed images employed in the single banknote recognition phase of the system drawn from the four datasets used namely Naira (First Row), Euro notes (Second Row), US Dollar (Third Row), and Canadian Dollar (Fourth Row). a) Comprises banknotes without clutters or occlusions captured with minimal illumination interference, b) Consists of Partially Occluded Banknotes with clutters and distortions, and c) Made up of notes with wrinkling and folds held by a user in an open environment with minutiae illumination obstruction.

While Figure 7 measures the performance accuracy using SIFT, SURF, HOG, GLOH, Shape Context, and Modified HOG using KNN, WKNN, RF, MLP, OVO-SVM classifiers. The x-axis shows the accuracy while the y-axis gives the extracted features and the z-axis has the classifiers for the Naira, USD, CAD, and Euro Banknotes. From the results, the best accuracy for the Features and classifiers were SURF/Modified HOG and OVO-SVM while the worst performance was achieved with SIFT and KNN Respectively although the proposed Modified HOG presented comparatively higher accuracy in most cases it was barely higher than the SURF techniques. Furthermore, RF and MLP did produce promising results, but their accuracy was dependent on the descriptor and currency as RF outperformed MLP in some cases and MLP outweighed RF in performance accuracy in other feature and classifiers combination. Although, WKNN outperformed KNN they both still had low rates. Lastly, the most inferior combination accuracy was SIFT and KNN with an average of 82%, and the highest accuracy combination was Modified HOG and OVO-SVM with an average rate of 97.3% with the proposed SURF closely following behind with an accuracy of 97.29%.

Further experiments were conducted with feature reduction techniques to remove features that hinder the accuracy and precision of the proposed approach, and the error rate was utilized in measuring the performance of which is presented in Table 3 above. The table gives the Error Rate of OVO-SVM of the different Descriptors and Feature Reduction on the Four Datasets.

The proposed approach Modified HOG/LVHCF presented the lowest error rate within the average rate of 0.0270 after feature reduction while shape context/PCA had the worse error rate with an average of 0.0585. The rankings of the descriptor and Feature reduction technique ranging from best to worse are Shape Context, SIFT, GLOH, HOG, SURF, Modified HOG and PCA, BOW, LVF, HCF, LVHCF respectively. Additionally, each feature reduction approach was applied to all the descriptors to determine the best pair with the highest recognition rate. From the result, PCA had the lowest acceptance rate across all data sets while Modified HOG/LVHCF presented the best results across all datasets with relatively higher performance than SURF. Regarding datasets, Naira notes offered the best result while CAD had the worse rate although there were cases where CAD outperformed Euro and USD data sets when SIFT, SURF and Shape Context descriptors were employed. Lastly, the proposed approach presented promising results in comparison with existing techniques and can be engaged in descriptor dimension reduction.

4.2. Counterfeit Detection (Experiment Two)

Figure 8 depicts genuine and fake banknotes employed in this research, the first row consists of Naira notes, the second row has Euro paper money, the third row with the Canadian dollar, and the fourth/last row has US dollar. The dataset covers a wide variety of conditions, such as occlusion, rotation, changes in scaling, illumination, and viewpoints.

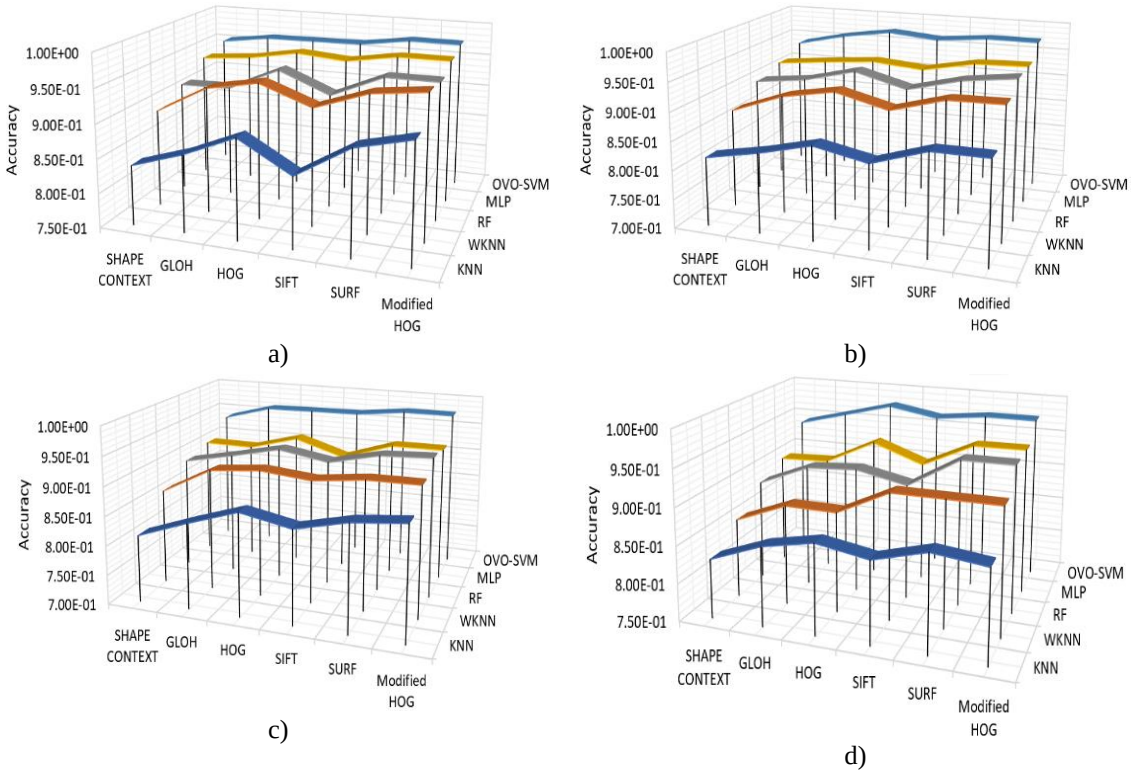


Figure 7: Accuracy measure of a) Naira, b) USD, c) CAD, and d) Euro using feature descriptors on different classifiers

The table presents the evaluation results of the proposed system on the data sets using OVO-SVM as it presented the best result during classification. For testing the system used 3110 samples for the Naira notes, 1943 instances for US Dollar, 1733 samples for Canadian Dollar and 2420 images of the Euro banknotes. Furthermore, the table also showed the Naira relatively had the highest correctly detected samples while the US dollar had the lowest successfully authenticated notes. However, all data sets had little incorrectly detected samples irrespective of the size of the data set; the system identified all images in the four datasets as paper money still, although some bills were wrongly classified. Finally, The Naira had the highest detection rate, the Euro came second, USA Dollar was third, and the Canadian dollar had the least authentication rate.



a) b)

Figure 8. a) Genuine and b) Fake Banknotes

4.3. Multiple Banknote Recognition (Experiment Three)

For the multiple banknote recognition systems, an object detection method in a cluttered scene using point feature matching of extracted HOG features based on image patches computed from the vertices of SURF points normalized using low variance and high correlation filter. It reads the test image and preprocesses it by applying grayscale conversion, noise removal, background subtraction, and normalization. Furthermore, the HOG Feature points are detected and used in matching the test image containing multiple banknotes is paired with the template image using the putative matching technique. The approach computes the transformation matched points while eliminating outliers. This aids in localizing the object in the scene. Lastly, the system locates the bounding polygon of the detected image and draws a black box to indicate the denomination.



a) b) c) d)

Figure 9. Multiple banknotes from a) Naira, b) CAD, c) Euro, and d) USD Dataset

Table. 3 Error Rate of feature reduction on different descriptors using OVO-SVM on the Four Datasets

No.	Descriptor	Feature Reduction	Naira	USD	CAD	Euro
		No Feature reduction	0.0489	0.0627	0.0630	0.0594
		PCA	0.1500	0.1948	0.2104	0.1956
		BOW	0.1300	0.1553	1.7267	0.1678
1	Shape Context	LVF	0.0950	0.1269	1.3056	0.1284
		HCF	0.0650	0.0754	0.0865	0.7253
		LVHCF	0.0700	0.0654	0.8721	0.6893
		No Feature reduction	0.0387	0.0421	0.0395	0.0415
		PCA	0.1300	0.1357	0.1243	0.1364
		BOW	0.0930	0.0892	0.7507	0.0885
2	GLOH	LVF	0.0600	0.0648	0.6340	0.0655
		HCF	0.0550	0.0540	0.5425	0.0574
		LVHCF	0.0575	0.0404	0.0409	0.0529
		No Feature reduction	0.0379	0.0283	0.0379	0.0233
		PCA	0.1000	0.1092	0.1150	0.1144
		BOW	0.0900	0.0921	0.0988	0.1077
3	HOG	LVF	0.0780	0.0795	0.0815	0.0921
		HCF	0.0550	0.0484	0.0502	0.0665
		LVHCF	0.0383	0.0285	0.0380	0.0238
		No Feature reduction	0.0371	0.0365	0.0368	0.0354
		PCA	0.1010	0.1204	0.1155	0.1288
4	SIFT	BOW	0.0750	0.0764	0.0843	0.0751
		LVF	0.0586	0.0582	0.0577	0.0549
		HCF	0.0400	0.0488	0.0422	0.0445
		LVHCF	0.0380	0.0469	0.0388	0.0363
		No Feature reduction	0.0261	0.0276	0.0268	0.0275
		PCA	0.0907	0.1048	0.1166	0.1287
		BOW	0.709	0.0977	0.0928	0.0845
5	SURF	LVF	0.558	0.0660	0.0702	0.0681
		HCF	0.350	0.0408	0.0502	0.0455
		LVHCF	0.0295	0.0389	0.3049	0.0315
		No Feature reduction	0.0266	0.0279	0.0268	0.0274
		PCA	0.0837	0.0884	0.0874	0.0865
		BOW	0.0649	0.0651	0.0662	0.0672
6	Modified HOG	LVF	0.0351	0.0257	0.0318	0.0328
		HCF	0.02660.	0.0384	0.0296	0.0299
		LVHCF	0265	0.0277	0.0268	0.0276

Figure 9 shows multiple banknote samples utilized in this research: a) comprises Naira notes, b) Canadain Dollar notes c) Euro bills, and d) US Dollar. The first row contains two bank notes; the second row contains three banknotes, the third row includes four bank notes, and the last/fourth row has five bank notes. The multiple notes dataset included clutters, distortions, rotations and overlapping captured in different illumination conditions.

While Figure 10 presents the database employed for the numerous banknote detection phases of the system, it consisted of 460 Naira images, 372 US dollar samples, 440 Canadian dollar images, and 365 Euro Samples. Table 6 shows the dataset of Naira, Euro, CAD and US notes used in the multiple banknote recognition experiments ranging from images with one banknote to five bills. It shows the precision and recall of this analysis; this was computed to investigate the effectiveness of our systems in distinguishing many banknotes in one image and determine which of the several banknote classes can best predict the note denomination irrespective of distortions.

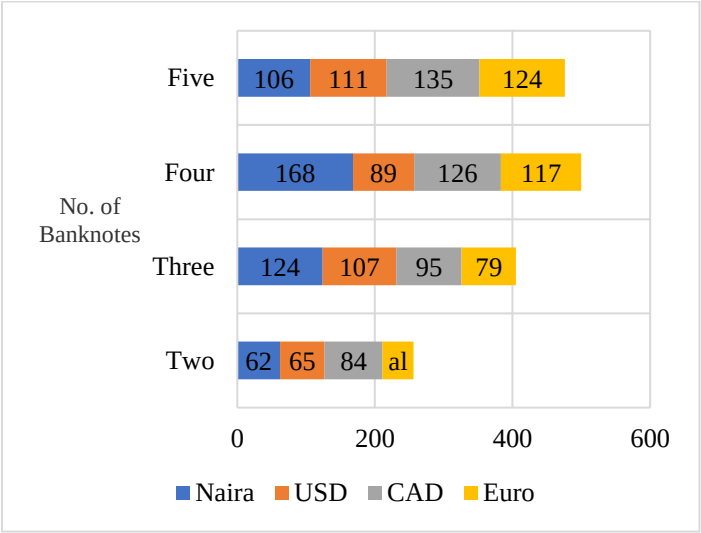


Fig. 10. Multiple banknotes Dataset

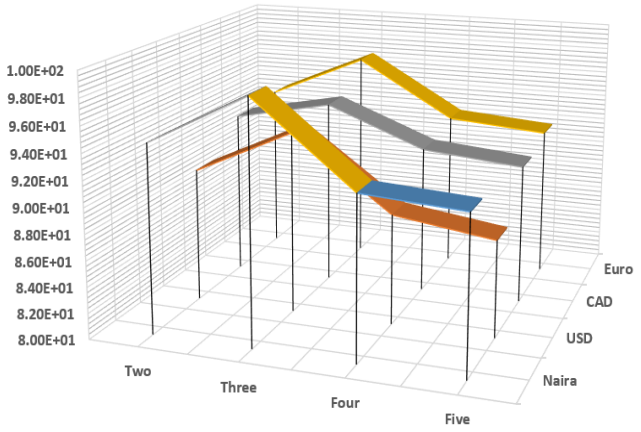


Figure 11. Recognition Rates of a) Naira b) USD c) CAD and d) Euro on multiple banknotes

Furthermore, in Figure 11 we have measured the Recognition Rate using One, two, three, four and five notes on the Naira, USD, CAD, and Euro Database. For simplicity, we have shown here only the recognition rate on the x-axis, while the number of the banknotes is fixed on the y-axis and the currency type on the z-axis. The reason for this experiment is to evaluate the effectiveness of the proposed feature extraction approach on multiple banknote recognition. The system recognized several currencies including same notes in the same image, separated, overlapped, cluttered and partially obstructed. After three bills the acceptance result reduced when the number of banknotes was increased. In all cases, the accuracy measure is more than 87%. Although one, two and three notes show more than 92% while four and five ranged between 87.50% and 92.45%. For this research, the best recognition rate across all datasets was three bills using Naira at 98.60%, and the worse was with five banknotes using USD at 87.50%.

Table 5. Qualitative Analysis of feature descriptors

Method	Time	Scale & Rotation	Illumination	Affine
Shape Context	Better	Better	Common	Good
SIFT	Common	Best	Common	Good
SURF	Best	Common	Common	Good
HOG	Common	Common	Best	best
GLOH	Good	Best	Common	Good
Modified HOG	Best	Better	Best	Best

In Table 5, we empirically compare the performance of all feature descriptors for multiple banknote identifications concerning its variants in five distinct scenarios: speed, scale, rotation, illumination, and affine invariance. The results of the experiment were analyzed qualitatively to give a general idea of the performance of each algorithm in distinct scenarios using four grades: Common, Good, Better, and Best. As shown in the table above, Shape context has significant advancements in speed, scale, affine, and rotation changes. But not illumination change. SIFT is slow; it is invariant to rotation and scale, maintains a stable degree of stability for affine transformation and a low threshold to illumination change. SURF is fast and has better performance than SIFT and GLOH, but it is not stable to the rotation and illumination changes. HOG is robust to rotation, illumination, positioning, and affine modifications but very slow and scale variant. GLOH has excellent performance in scale and rotation, presented reasonable speed and affine changes and had a deficient sense of illumination invariance. The Modified HOG descriptor maintains a proper stability illumination, rotation, and affine transformation. It can also handle reasonable scale invariance and compared with the other algorithms; it has distinct advantages regarding speed. Thus, can be deduced that the proposed technique presented a promising performance in banknote recognition. Lastly, the result of this qualitative analysis is not constant in all cases as changes to the dataset, and classifiers can lead to a different conclusion.

4.4. Training Sample Needed

One critical issue is determining the number of notes by denomination and authenticity is needed to ensure a reasonable performance from the proposed technique. To solve this problem, the various databases are evaluated concerning the sizes by varying the training set dimension. A remarkable increment on the recognition and error rate is seen when the training data set increases from 300 to 800 for banknote recognition, 100 to 250 for counterfeit detection, and 50 to 200 for multiple currency recognition. After which growing the data further only produced asymptotical performance improvement but had no effect on its accuracy and error rate, and this could be attributed to the inclusion of more distorted notes.

5. Conclusion and Future Work

This research was conducted to determine the efficiency of selected Modified HOG/LVHCF descriptor for automatic recognition of paper money. The proposed system was implemented using the Nigerian notes, and for experimental evaluation of the system, a further test was conducted on USD, CAD, and Euro notes. The proposed method extracted the HOG descriptor from image patches created from the vertices of SURF interest points feature of the banknotes, Low Variance, and High Correlation feature reduction filter was employed to reduce the feature set and classified using OVO-SVM in identifying and authenticating fake and genuine bills. Lastly, point matching technique was applied to detect feature vectors to match images containing multiple banknotes to determine the value of the currencies. Future research would include incorporating new feature descriptors and reduction approach to reduce processing time in addition to scaling down the application to work on mobile devices and reducing the computational complexity to enable embedding in micro-controllers and microchips.

Acknowledgment

This work was supported partly by the Centre for Pattern Recognition and Machine Intelligence (CENPARMI) Research Group of Concordia University and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] D. Alekhya, G. Devi Surya Prabha, G. Venkata Durga Rao, "Fake currency detection using image processing and other standard methods," IJRCCT, 3(1), pp. 128-131, 2014.

[2] M. Thakur, A. Kaur, "Various Fake Currency Detection Techniques," International Journal for Technological Research in Engineering, 1(11), 2014.

[3] A. Berenguel, O.R. Terrades, J. Lladós, C. Cañero, "Banknote counterfeit detection through background texture printing analysis," 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 66-71, 2016.
[4] M.S. Uddin, P.P. Das, M.S.A. Roney, "Image-based approach for the detection of counterfeit banknotes of Bangladesh," (ICIEV), 5th International Conference on Informatics, Electronics and Vision, pp. 1067-1072, 2016.
[5] H. Hassanpour, A. Yaseri, G. Ardeshiri, "Feature extraction for paper currency recognition," ISSPA 2007. 9th International Symposium on Signal Processing and Its Applications, pp. 1-4, 2007.
[6] N. Kato, S. Omachi, "A handwriting character recognition system using directional element feature," IEEE Trans. Pattern Anal. Mach. Intell, 21(3), pp. 258–262, 1999.
[7] F. García-Lamont, J. Cervantes, A. López, "Recognition of Mexican banknotes via their color and texture features," Expert Systems with Applications, 39(10), pp. 9651-9660, 2012.
[8] H. Aggarwal, P. Kumar, "Indian Currency Note Denomination Recognition in Color Images," International Journal on Advanced Computer Engineering and Communication Technology, 1(1), pp. 12-18, 2014.
[9] P.D. Pawar, S.B. Kale, "Recognition of Indian Currency Note Based on HSV Parameters," International Journal of Science and Research (IJSR), 3(6), pp. 132-137, 2014.
[10] J. Lee, J. Seong-Goo, K. II-Hwan, "Distinctive point extraction and recognition algorithm for various kinds of euro banknotes," International Journal of Control, Automation, and Systems, 2(2), pp. 201-206, 2004.
[11] M. Ionescu, A. Ralescu, "Fuzzy hamming distance-based banknote validator," FUZZ'05, The 14th IEEE International Conference on Fuzzy Systems, pp. 300-305, 2005.
[12] H. Hassanpour, M.F. Payam, "Using Hidden Markov Models for paper currency recognition," Expert Systems with Applications, 36(6), pp. 10105-10111, 2009.
[13] F. Poor, T. Mohammadpour, A. Kianis, "Persian Banknote Recognition Using Wavelet and Neural Network [A]," in Proceedings of the IEEE Conference on Computer Science and Electronic Engineering, 3, pp. 679-684, 2012.
[14] W. Selesnick, R.G. Baraniuk, N.G. Kingsbury, "The dual-tree complex wavelet transform," IEEE Trans. Signal Process, 22(6), pp. 123–151, 2005.
[15] W.L. Chan, H. Choi, and R.G. Baraniuk, "Coherent multiscale image processing using dual-tree quaternion wavelets," IEEE Transactions on Image Processing, 17(7), pp. 1069-1082, 2008.
[16] S. Gai, G. Yang, M. Wan, "Employing quaternion wavelet transform for banknote classification," Neurocomputing, 118, pp. 171-178, 2013.
[17] F. Ahangaryan, T. Mohammadpour, A. Kianisarkaleh, "Persian Banknote Recognition Using Wavelet and Neural Network," (ICCSEE), 2012 International Conference on Computer Science and Electronics Engineering, 3, pp. 679-684, 2012.
[18] A. Rajaei, E. Dallalzadeh, M. Imran, "Feature extraction of currency notes: an approach based on wavelet transform. (ACCT)," Second International Conference on Advanced Computing & Communication Technologies, pp. 255-258, 2012.
[19] C. Harris, M. Stephens, "A combined corner and edge detector" In Alvey vision conference, 15(50), pp. 147-151, 1988.
[20] T. Lindeberg, "Feature detection with automatic scale selection," International Journal of computer vision, 30(2), pp. 79–116, 1998.
[21] K. Mikolajczyk, C. Schmid, "Indexing based on scale invariant interest points," ICCV Proceedings. Eighth IEEE International Conference on Computer Vision, 1, pp. 525–531, 2001.
[22] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of computer vision, 60(2), pp. 91–110, 2004.
[23] T. Reiff, P. Sincak, "Multi-agent sophisticated system for intelligent technologies," ICCC International Conference on Computational Cybernetics, pp. 37-40, 2008.
[24] K. Van De Sande, T. Gevers, C. Snoek, "Evaluating color descriptors for object and scene recognition," IEEE transactions on pattern analysis and machine intelligence, 32(9), pp. 1582-1596, 2010.
[25] T.H. Rassem, B. E. Khoo, "Object class recognition using combination of color SIFT descriptors," IEEE International Conference on Imaging Systems and Techniques (IST), pp. 290-295, 2011.
[26] N.S. Pai, J.H. Hong, P.Y. Chen, J.K. Wu, "Application of design of image tracking by combining SURF and TLD and SVM-based posture recognition system in robbery pre-alert system," Multimedia Tools and Applications, 76(23), pp. 25321-25342, 2017.
[27] F.M. Hasanuzzaman, Y. Xiaodong, T. Yingli, "Robust and effective component-based banknote recognition for the blind," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), pp. 1021-1030, 2012.
[28] W. Chen, S. Ding, Z. Chai, D. He, W. Zhang, G. Zhang, W. Luo, "FPGA-Based Parallel Implementation of SURF Algorithm," (ICPADS). 22nd International Conference on Parallel and Distributed Systems, pp. 308-315, 2016.
[29] Z.R. Wang, Y.L. Jia, H. Huang, S.M. Tang, "Pedestrian detection using boosted HOG features," ITSC 11th International IEEE Conference on Intelligent Transportation Systems, pp. 1155-1160, 2008.
[30] T.V. Dittimi, A.K. Hmood, C.Y. Suen, "Multi-class SVM based gradient feature for banknote recognition," (ICIT), IEEE International Conference on Industrial Technology, pp. 1030-1035, 2017.
[31] N.K. Verma, R. Abhishek, S. Al, "An optimized fault diagnosis method for reciprocating air compressors based on SVM," (ICSET), IEEE International Conference on System Engineering and Technology, pp 65-69, 2011.
[32] A. Rashid, A. Prati, R. Cucchiara, "On the design of embedded solutions to banknote recognition," Optical Engineering, 52(9), pp. 093106-093106, 2013.
[33] A. Govada, G. Bhavul, K.S. Sanjay, "Distributed multi-class SVM for large datasets," In Proceedings of the Third International Symposium on Women in Computing and Informatics, pp. 54-58, 2015.
[34] T.V. Dittimi, C.Y. Suen, "Mobile App for Detection of Counterfeit Banknotes," In Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, 31, pp. 156-168, 2018.
[35] T. Dittimi, C. Suen, A. Hmood, "Mobile Based Assistive Technologies for Authentication of Banknotes," ICPRAI 2018 - International Conference on Pattern Recognition and Artificial Intelligence, pp. 366-371, 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 363-371 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Amplitude-Frequency Analysis of Emotional Speech Using Transfer Learning and Classification of Spectrogram Images

Margaret Lech[*,1], Melissa Stolar[1], Robert Bolia[2], Michael Skinner[2]

[1]*School of Engineering, RMIT University, VIC 3000, Australia*

[2]*Defence Science and Technology Group, VIC 3207, Australia*

ARTICLE INFO

Article history:
Received: 06 June, 2018
Accepted: 03 August, 2018
Online: 25 August, 2018

Keywords:
Speech processing
Emotion recognition
Deep neural networks

ABSTRACT

Automatic speech emotion recognition (SER) techniques based on acoustic analysis show high confusion between certain emotional categories. This study used an indirect approach to provide insights into the amplitude-frequency characteristics of different emotions in order to support the development of future, more efficiently differentiating SER methods. The analysis was carried out by transforming short 1-second blocks of speech into RGB or grey-scale images of spectrograms. The images were used to fine-tune a pre-trained image classification network to recognize emotions. Spectrogram representation on four different frequency scales - linear, melodic, equivalent rectangular bandwidth (ERB), and logarithmic - allowed observation of the effects of high, mid-high, mid-low and low frequency characteristics of speech, respectively. Whereas the use of either red (R), green (G) or blue (B) components of RGB images showed the importance of speech components with high, mid and low amplitude levels, respectively. Experiments conducted on the Berlin emotional speech (EMO-DB) data revealed the relative positions of seven emotional categories (anger, boredom, disgust, fear, joy, neutral and sadness) on the amplitude-frequency plane.

1. Introduction

This paper is an extension of work originally presented in the 11th International Conference on Signal Processing and Communication Systems, ICSPCS'2017 [1]. The conference work introduced a new, efficient, real-time speech emotion recognition (SER) methodology. A demo of this application can be found on: (https://www.youtube.com/watch?v=cIsVGiFNJfE&t=41).

In this study, the method proposed in [1] was used for analytical purposes to determine how different emotional categories are coded into the amplitude-frequency characteristics of emotional speech. The underlying assumption was that, the accuracy of detecting a particular emotion depends on how many cues related to this emotion are given by the input speech features. By selecting acoustic speech features representing or emphasizing different frequency and amplitude values, and observing which emotions were best recognised, the study determined links between different emotions and corresponding amplitude-frequency characteristics of speech. In other words, features that lead to higher classification accuracy for a given emotion were thought to be highly representative of this emotion.

The search for the "best" or the most representative acoustic features has been one of the most important themes driving SER, as well as emotional speech synthesis studies [2]. In SER, this knowledge can improve differentiation between emotions, and in speech synthesis, it can lead to more natural sounding emotional speech. Despite extensive research, it is still unclear what the "best" features are, as it is uncertain how different emotions are coded into the time-varying amplitude-frequency characteristics of speech.

The remaining sections of the paper are organised as follows. Section 2 gives a brief review of related studies. Section 3 explains the methodology. The experiments and discussion of results are presented in Section 4, and Section 5 provides the conclusion.

2. Previous Works

Traditionally, the links between speech acoustics and emotions have been investigated by classifying emotional speech samples

[*]Corresponding Author: Margaret Lech, School of Engineering, RMIT University, GPO Box 2476, 3000 VIC, Australia, Contact No: +61 3 99251028, Email: margaret.lech@rmit.edu.au

based on various low-level parameters, or groups of parameters, defined by models of speech production. Examples of low-level parameters include the fundamental frequency (F0) of the glottal wave, formant frequencies of the vocal filter, spectral energy of speech, and speech rate. The low-level features were later enriched by the addition of higher level derivatives and statistical functionals of the low-level parameters. The Munich Versatile and Fast Open-Source Audio Feature Extractor (openSMILE) [3] gives the current software standard, allowing for the calculation of over 6000 low- and high-level acoustic descriptors of speech. In [4] links between acoustic properties of speech and emotional processes in speakers performing a lexical decision task were investigated. Basic parameters such as F0, jitter and shimmer were found to be correlated with emotional intensity. The results indicated that acoustic properties of speech can be used to index emotional processes, and that characteristic differences in emotional intensity may modulate vocal expression of emotion. However, acoustic parameters characterizing different types of emotions were not investigated.

It has been often assumed, especially in earlier studies, that unlike facial expressions that convey qualitative differences between emotions (emotional valence), speech can only communicate emotional intensity (arousal) [5,6]. This assumption was shown to be incorrect by proving that judges are almost as accurate in inferring different emotions from speech as from facial expression [7,8]. As suggested in [6], it is possible that F0, energy, and rate may be most indicative of arousal while qualitative, valence differences may have a stronger impact on source and articulation characteristics. One of the tables presented in [8] showed how selected acoustic speech parameters vary with different emotional categories. Namely, speech intensity, mean, variability and range of F0, sentence contours, high-frequency energy, and articulation rate have been compared across six emotional states (stress, anger/rage, fear/panic, sadness, joy/elation, and boredom). An investigation described in [9] shows links between emotions and prosody features including intonation, speaking rate and signal energy. Correlation between naturally expressed anger and despondency, and various acoustic cues of the speech signal of 64 call centre customers was analysed in [10]. It was found that anger was characterized by a rise of F0 and speech amplitude, whereas despondency by a decreased F0 and syllable rate. Teager energy operator (TEO) parameters [11-13], which simultaneously estimate the instantaneous energy and frequency changes of speech signals, have been shown to be effective in the classification of speech under stress [14,15] and in automatic SER [16]. It was suggested that the high effectiveness of these parameters is due to their sensitivity to the acoustic effects caused by a nonlinear air flow and the formation of vortices in the vicinity of the vocal folds. A discussion of the links between emotions and some of the basic acoustic features including F0, formants, vocal tract cross-section, mel-frequency cepstral coefficients (MFCCs) and TEO features, as well as the intensity of the speech signal and the speech rate can be found in [13]. The high velocity of air moving through the glottis and causing vocal fold vibration was found to be indicative for music-like speech such as joy or surprise while, low velocity was an indicator of harsher styles such as anger or disgust [17]. Identification of the "best" frequency sub-bands of the power spectrum for the classification of different emotions has not been conclusive. A number of studies [18-20] have indicated high importance of the low frequency band ranging from 0 to 1.5 kHz for speech emotions, whereas [21] pointed to the high frequency range (above 1.5 kHz). In He [16] discriminative powers of different frequency bands were examined. It was found that the largest diversity between energy contributions from different emotions occurs in the low frequency range of 0-250 Hz and the high frequency range of 2.5-4 kHz. The middle range of 250 Hz to 2.5 kHz did not show clear differences between emotions. Some of the most often used parameters in speech classification across different databases and languages are the MFCCs [22]. In general, MFCCs are thought to give good representation of the speech signal by considering frequency response characteristics of the human auditory system. However, applications of MFCCs into the SER task have shown relatively poor results [14,16,21,23]. The log-frequency power coefficients (LFPCs) calculated within critical bands of the human auditory system [24] were found to outperform the MFCCs [21]. Experiments presented in [16] examined and ranked 68 different low and high-level features for emotion recognition from natural (non-acted) speech. TEO parameters estimated within perceptual wavelet packet frequency bands (TEO-PWP), as well as the area under the speech energy envelope (AUSEE), have been found to produce the highest performance in both natural emotion and stress recognition.

As shown in the above few examples of the very rich research field, the task of finding links between categorical emotions and acoustic speech parameters has been particularly challenging due to inconsistencies between studies, which include the ways emotions are evoked in speakers (natural, acted, induced), types of emotional labels (different emotional categories or arousal/valence) and types of acoustic speech parameters [8,12,25]. To deal with these difficulties, current trends in automatic speech emotion recognition (SER) are shifting away from learning differences between individual low- or high-level parameters, and moving towards using speech spectrograms that capture the entire time-evolution of emotional acoustics in the form of two-dimensional, time-frequency, spectral magnitude arrays.

Deep neural network structures have been trained to efficiently classify speech spectrograms to recognize different emotional categories [26][33]. Although this methodology is very powerful, the current state-of-the-art results indicate that there is still room for improvement. In particular, the inter-emotional confusion tables show high levels of misclassification between neutral and dysphoric [16], neutral and boredom [34], and joy and anger [16],[34]. Systems designed with prior knowledge of emotional cues have been shown in the past to be particularly effective in increasing the overall SER accuracy. In [35][36] for example, a saliency analysis was applied to speech spectrograms to determine regions of the time-frequency plane that provide the most important emotional cues.

Application of these regions as inputs to a Convolutional Neural Network (CNN) led to high SER accuracy, however, the saliency analysis had a time-varying character and involved very high computational cost. This study offers a more computationally efficient approach to finding what parts of the amplitude-frequency plane carry the most important cues for different emotions. It is expected that the knowledge of the amplitude-frequency

characteristics of individual emotions will facilitate future designs of more efficient SER systems with reduced inter-class confusion.

Table 1: Speech and image data description.

Emotion	No. of speech samples (utterances)	Total duration of speech samples [sec]	No. of generated spectrogram images (RGB or Grey-Scale)
Anger	129	335	27220
Boredom	79	220	18125
Disgust	38	127	11010
Fear	55	123	5463
Joy	58	152	12400
Neutral	78	184	14590
Sadness	53	210	18455
TOTAL	390	1207	111425

3. Method

3.1. Speech Data

The study was based on the EMO-DB database [37] that is routinely used in evaluations of SER systems. It contains speech recordings collected from 10 professional actors (5 male and 5 female) speaking in fluent German. Each actor simulated 7 emotions (anger, joy, sadness, fear, disgust, boredom and neutral speech) while pronouncing 10 different fixed-text utterances consisting of single sentences with linguistically neutral contents. Each actor pronounced each utterance with a different emotion, however in some cases the speakers provided more than one version of the same utterance. Table 1 shows the numbers and total duration of available speech samples (pronounced utterances) for each emotion. The original recordings were validated using listening tests conducted by 10 assessors; details can be found in [37]. Only speech samples that scored recognition rates greater than 80% were used in this study. The sampling rate was 16 kHz, giving 8 kHz speech bandwidth. Despite more recent developments of emotional speech datasets, the EMO-DB remains one of the best and most widely used standards for testing and evaluating SER systems. The important strength of the EMO-DB is that it offers a good representation of gender and emotional classes. The main disadvantage is that the emotions appear to be acted in a strong way, which in some cases may be considered as unnatural.

3.2. SER Using Transfer Learning and Images of Spectrograms

Existing state-of-the-art SER methods apply deep Convolutional Neural Networks (CNNs) trained on spectral magnitude arrays of speech spectrograms [26-29]. To achieve high accuracy, complex CNN structures must be trained on a very large number (in the order of millions) of labelled spectrograms. This method (known as "fresh training") is computationally intense, time consuming and requires large graphic processing units (GPUs).

At present the availability of large datasets of emotionally labelled speech is limited. On the other hand, in many cases, close to the state-of-the-art results can be achieved using a much simpler process of transfer learning. In transfer learning a small, problem-specific data set is used to fine-tune an existing network that has been already pre-trained on a very large amount of more general data.

Table 2: Fine-tuning parameters for AlexNet (Matlab version 2017b)

Parameter	Value
Minibatch size	128
Maximum number of epochs	5
Weight decay	0.0001
Initial learning rate	0.0001
Weight learn rate factor	20
Bias learn rate factor	20

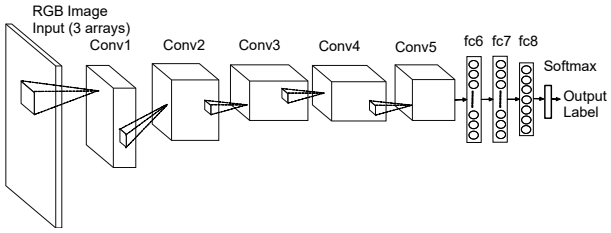


Figure 1. Structure of AlexNet. The input consists of 3 arrays (3 input channels). Feature extraction is performed by 5 convolutional layers (Conv1-5), and classification by 3 fully-connected layers (fc5-8). The output is given in the form of soft labels indicating probability of each class.

The current study applied transfer learning to a pre-trained, general-purpose image classification network known as AlexNet [38], with inputs given as images (not magnitude arrays) of speech spectrograms. AlexNet has been trained on over 1.2 million images from the ImageNet [39] data representing 1000 classes. As shown in Figure 1, AlexNet is a CNN [40] that consists of a 3-channel input layer followed by five convolutional layers (Conv1-Conv5) along with max-pooling and normalization layers, and three fully connected layers (fc6-fc8). The output from the last layer is passed through the normalized exponential Softmax function [41] that maps a vector of real values (that sum to 1) into the range [0, 1]. These values represent the probabilities of each object class. In the experiments described here, the final classification label was given by the most probable class (emotion). The fine tuning of AlexNet was performed within the Matlab (version 2017b) programming framework [42]. The network was optimized using stochastic gradient descent with momentum (SGDM) and L2 regularization factor applied to minimize the cross-entropy loss function. Table 2 provides values of the tuning parameters. Given this choice of parameters, the fine-tuning process changed mostly the final, fully connected (data-dependent) layers of the network leaving the initial (data-independent) layers almost intact.

Despite the fact that in recent years AlexNet has been rivalled by newer, significantly more complex and more data demanding networks, the pre-trained structure is still of great value, as it provides a good compromise between data requirements, network simplicity and high performance. Informal tests on more complex networks such as the residual network ResNet-50, the Oxford Visual Geometry Group networks VGG-16 and VGG-19, and the GoogleLeNet have shown that for a given outcome, the training time needed by the larger networks was significantly longer than for AlexNet. It could also imply that a larger dataset may be needed, but it would have to be confirmed by formal testing.

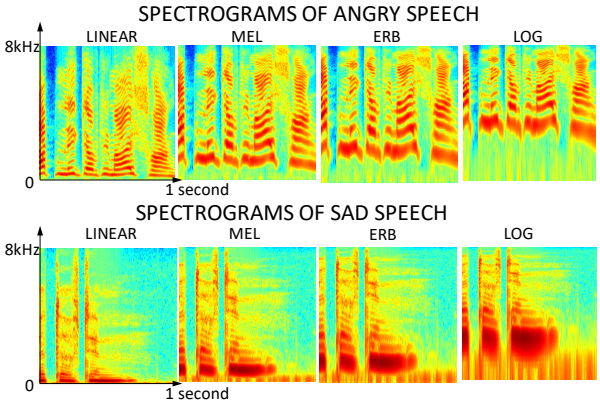


Figure 2. Examples of RGB images of speech spectrograms for the same sentence pronounced with sadness and anger, and depicted on four different frequency scales linear, mel, ERB and log. The linear scale represents the high frequency details, mel scale-the mid-high frequency details, ERB – the mid-low frequency details and log – the low frequency details.

3.3. Generation of Spectrogram Images

The amplitude-frequency analysis of the characteristics of emotional speech was achieved by analyzing the SER results given by different forms of features (i.e. spectrogram images representing speech signals. The SER was performed by applying transfer learning to the pre-trained AlexNet with input arrays given as RGB or grey-scale images depicting speech spectrograms [1] (see Figure 4). To preserve the feasibility of real-time SER [1,26,27], the features were generated on a frame-by-frame basis, and no utterance based parameters were calculated. Since the emotional labels for the EMO-DB data were given for entire utterances, the emotional label for each frame was assumed to be the same as the label of the utterance from which it was extracted. Short-time Fourier transform spectrograms were computed for 1-second blocks of speech waveforms using Hamming window frames of length 25 milliseconds with 12.5 milliseconds of overlap between frames. Short silence intervals occurring within sentences were kept intact. The calculations were performed using the Matlab Voicebox *spgrambw* procedure [43], which generated magnitude spectrogram arrays of size 257 x 259 for the frequency and time dimensions, respectively. The stride between subsequent 1-second blocks was 10 milliseconds. Table 1 shows the numbers of images generated from the available speech data.

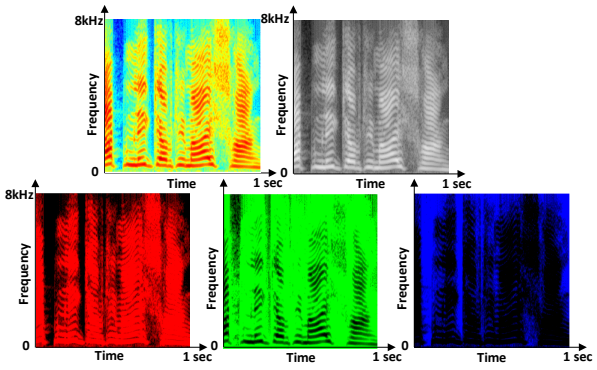


Figure 3. An example of the RGB, grey-scale, R, G and B spectrogram images for the same speech sample of emotionally neutral speech. The R components emphasizes high-amplitude details, the G component – the mid amplitude details and the B component – the low-amplitude details.

The experiments investigated and compared four frequency scales of speech spectrograms: linear, melodic (mel) [44], equivalent rectangular bandwidth (ERB) [45] and logarithmic (log) [46]. While the alternative frequency scales were applied along the vertical axis of the spectrograms from 0 to 8 kHz, the horizontal time scale was in all cases linear spanning the time range of 0 to 1 second. The dynamic range of spectral magnitudes was normalized from -130 dB to -22 dB, representing the respective minimum (Min) and maximum (Max) magnitude values. These values were calculated for magnitude spectrograms over the entire database. The magnitude spectrograms were transformed into RGB or grey-scale images. The RGB images were generated with the Matlab "jet" colormap [47,48] containing 64 default colors. After downsampling to 227 x 227 pixels (using the Matlab *imresize* command), the spectrogram images provided input arrays to AlexNet (Figure 4). More details on the AlexNet input arrangements are given in Section 4.

3.4. Frequency Representation by Different Scales of Spectrogram Images

Four different frequency scales (linear, mel, ERB and log) [49] were applied when generating the spectrogram images in order to visually emphasize different frequency ranges. Figure 2 shows examples of spectrograms for the same sentence pronounced with sad and angry emotion plotted on four different frequency scales: linear, mel, ERB and log [49]. This order of scales corresponds to the process of gradually "zooming into" the lower frequency range features (about 0 to 2 kHz), and at the same time "zooming out" of the higher frequency range features (about 2 kHz to 8 kHz) features. Therefore, the application of different frequency scales effectively provided the network with either more- or less-detailed information about the lower or upper range of the frequency spectrum.

Thus, the linear scale emphasized details of the high-frequency (6-8 kHz) components of speech characterizing unvoiced consonants. The mel and ERB scales provided details of the mid-high (4-6 kHz) and mid-low (2-4 kHz) ranges, respectively (characteristic to both voiced and unvoiced speech), and the log scale provided the most details of the low-frequency (0.02-2 kHz) range (important for vowels and voiced consonants).

3.5. Amplitude Representation by R, G and B Images of Spectrograms

While the RGB images of spectrograms used in the speech classification gave visual representations of the time-frequency decomposition of speech signals, each of the R, G and B color components emphasized a different range of speech spectral amplitude values. As shown in Figure 3, the R component gave greatest intensity of the red color for high spectral amplitude levels, thus emphasizing details of the high-amplitude spectral components of speech (e.g. vowels and voiced consonants). The B component gave greatest intensity of the blue color for lower amplitudes, and therefore it emphasized details of the low-amplitude spectral components of speech (e.g. unvoiced consonants) and gaps between speech. Similarly, the G component emphasized details of the mid-range spectral amplitude components (both voiced and unvoiced).

3.6. Experimental Setup

A 5-fold cross-validation technique was used with 80% of the data distributed for the fine tuning, and 20% for the testing of AlexNet. All experiments were speaker- and gender-independent. The experimental framework is illustrated in Figure 4. Three different experiments, each having different data arrangements for the three input channels of the AlexNet were conducted as follows:

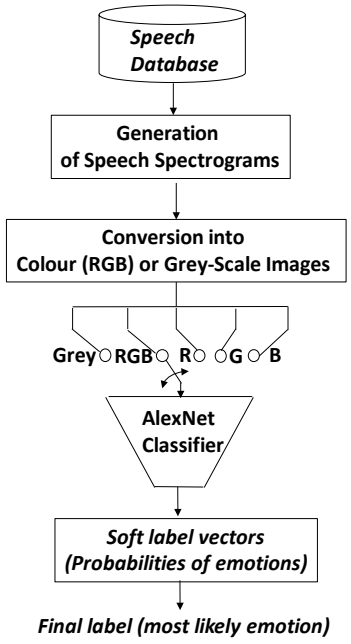


Figure 4. Framework diagram for the amplitude-frequency analysis of emotional speech.

- Experiment 1. To investigate the performance of the color RGB images of spectrograms on average and across emotions.

  AlexNet input: A different color-component array (R, G and B) was given as an input to each of the 3 input channels. The same process was repeated for four different frequency scales of spectrograms.

- Experiment 2. To investigate how the grey-scale image representation of spectrograms performs on average and across emotions.

  AlexNet input: Identical copies of grey-scale image arrays were given as inputs to each of the 3 input channels. It was repeated for four different frequency scales of spectrograms.

- Experiment 3. To investigate how each of the individual color components (R, G or B) performs on average and across emotions.

  AlexNet input: Identical copies of arrays representing the same color component were provided as inputs to each of the 3 input channels. It was repeated for four different frequency scales of spectrograms.

In contrast with the common practice of duplicating the information provided to the network channels [29], the first experiment provided more meaningful and complimentary information to each of the three processing channels of the neural network.

3.7. SER Performance Measures

The emotion classification performance was assessed using standard measures applied in SER tasks. It included accuracy, F-score, precision and recall [50][51] given by (1)-(4) respectively, and calculated separately for each emotion.

$$Accuracy = \frac{t_p+t_n}{t_p+t_n+f_p+f_n} \quad (1)$$

$$F-Score = 2\frac{Precision\ x\ Recall}{Precision+Recall} \quad (2)$$

$$Precision = \frac{t_p}{t_p+f_p} \quad (3)$$

$$Recall = \frac{t_p}{t_p+f_n} \quad (4)$$

Where, t_p and t_n denoted numbers of true positive and true negative classification outcomes, while f_p and f_n denoted numbers of false positive and false negative classification outcomes, respectively.

To reflect the fact that the emotional classes were unbalanced, the weighted average accuracy and F-score were estimated as:

$$Weighted\ Average = \frac{Q_{c_1}|c_1|+\cdots+Q_{c_N}|c_N|}{|c_1|+|c_2|+\ \ldots+\ |c_N|} \quad (5)$$

Where Q_{c_i} denoted either accuracy or F-score for the ith class ($i = 1,2,\ldots,N$) given as (1) or (2) respectively. The values of $|c_i|$ denoted class sizes and N was the number of classes.

4. Results and Discussion

The outcomes of the three SER experiments showed what type of input to AlexNet provided the most efficient training and emotion recognition results. By doing so, indirect insights into the spectral amplitude-frequency characteristics of different emotions were gained. It was achieved through both the application of different frequency scales of spectrograms that visually emphasized details of different frequency ranges, as well as the use of different color components that emphasized different spectral amplitude ranges of speech.

4.1. Comparison between Classification Performance of RGB, Grey-Scale and R, G and B Components

On average, the RGB images outperformed the individual color components and grey-scale images, giving the highest average classification accuracy and F-scores for all frequency scales (Figure 5). Both the RGB and grey-scale images achieved highest performance with the mel frequency scale. The largest difference between RGB and grey-scale images was observed for the ERB and log scales (about 2%). For the other scales, the difference was only about 1% showing that the high frequency details depicted by the ERB and log scales were likely to be clearly depicted by the color RGB images, and to some extend blurred by the grey-scale images.

In addition, Figure 5 shows that the best performing single-color component was green (G) followed by red (R), and the worse performing was blue (B). It indicates that the mid-amplitude

spectral components of speech are likely to be the most important for the differentiation between emotions. The low-amplitude components play a less important role, since they predominantly represent unvoiced speech and gaps of silence between words. For the R color-component, which emphasized high-amplitude speech components, the highest performance was given by the mel scale, which provided a detailed view of the mid-high frequency range. Therefore, the high-amplitude emotional cues appeared to be linked mostly to the mid-high frequency speech components (which amongst other things contain information about the harmonics of the glottal wave fundamental frequency (F0) and the higher formants of the vocal tract).

For the G color-component, the ERB scale, which reveals the mid-low frequency details of speech, was the best performing. This shows that the mid-amplitude emotional cues could be linked to the mid-low-frequency speech components (low harmonics of F0 and low formants). Finally, for the B color-component, the linear scale provided the best performance. This indicates that the low-amplitude emotional cues are predominantly linked to the higher frequency components of speech, such as unvoiced consonants, higher formants and higher harmonics of F0, as well as additional harmonic components generated due to the nonlinear air flow and vortex formation in the vicinity of the glottal folds [14].

4.2. Comparison between RGB, Grey-Scale and R, G and B Components across Different Emotions and Frequency Scales.

To analyze the performance of different spectrogram representations across different emotions, precision-recall graphs shown in Figures 6 and 7 were made. When using this representation, the classification aim was to achieve results that were close to the diagonal line of equal precision and recall values, and as close as possible to the top right corner showing the maximum values of these two parameters. Taking this into account, anger was detected with the highest precision/recall scores by the R-components using the ERB-scale, indicating that the anger cues are most likely to be encoded into the high amplitude and medium-low frequency components of speech.

Disgust was most efficiently detected by the RGB images using the logarithmic frequency scale, which points to the low frequency speech components spanning all amplitude values as carriers of this emotion. Joy was coded the same way as anger into the medium-low frequency and high amplitude components, as it was best detected by the R-image components and the ERB scale.

Boredom showed a distinctly different pattern by being best detected by the mel-scale and G-images. This means that boredom was coded into the medium-high frequencies and medium range amplitudes of speech.

Fear was also distinctly different by being best detected by the mel-frequency scale and grey-scale images. This points to the medium-high frequency range and all amplitude values of speech as carriers of this emotion.

Sadness gave the best performance for the mel-scale and RGB images, showing that like fear, it was coded into mid-high frequencies and all amplitude values of speech.

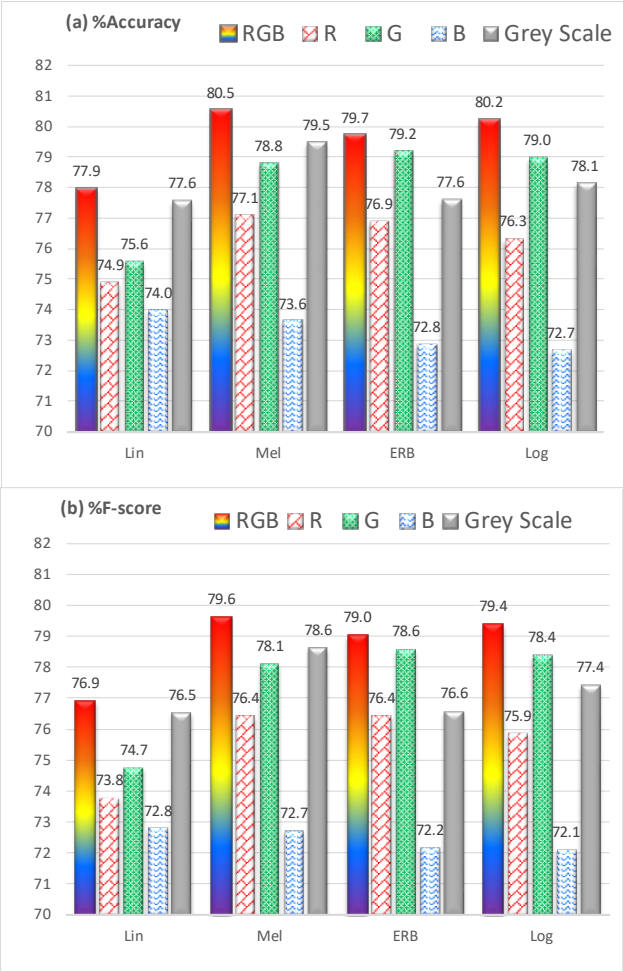


Figure 5. Average weighted classification accuracy (a) and F-score (b) using different frequency scales.

Finally, the emotionally neutral speech was best detected by the ERB scale and RGB images. Therefore, it was likely to be coded into the medium-low frequencies across all amplitudes values of speech. These observations are summarized in Table 3.

4.3. Statistical Significance Analysis

As shown in the above examples, different forms of spectrogram image representation lead to different classification outcomes. To find out which of the observed differences were statistically significant, one-way ANOVA analysis with Bonferroni correction was conducted using the SPSS package. The results comparing the average accuracy values obtained during SER are presented in Table 4. In 5 out of 10 possible pairs of spectrogram image representation, the differences were found to be statistically significant. Namely, the B components were significantly different compared to all other types of images, the R components were different compared to the RGB and B components, and finally the G components were significantly different only when compared to the B components. On average, there was no significant difference between the RGB images and G image-components. The least significant differences were obtained for the grey-scale images compared to the G image-components. This means that, in terms of the classification accuracy, the grey-scale representation appears to be very close to the G images - both

giving good visual information of the mid-amplitude components of speech, which, as previously discussed, were the most important carriers of emotional cues.

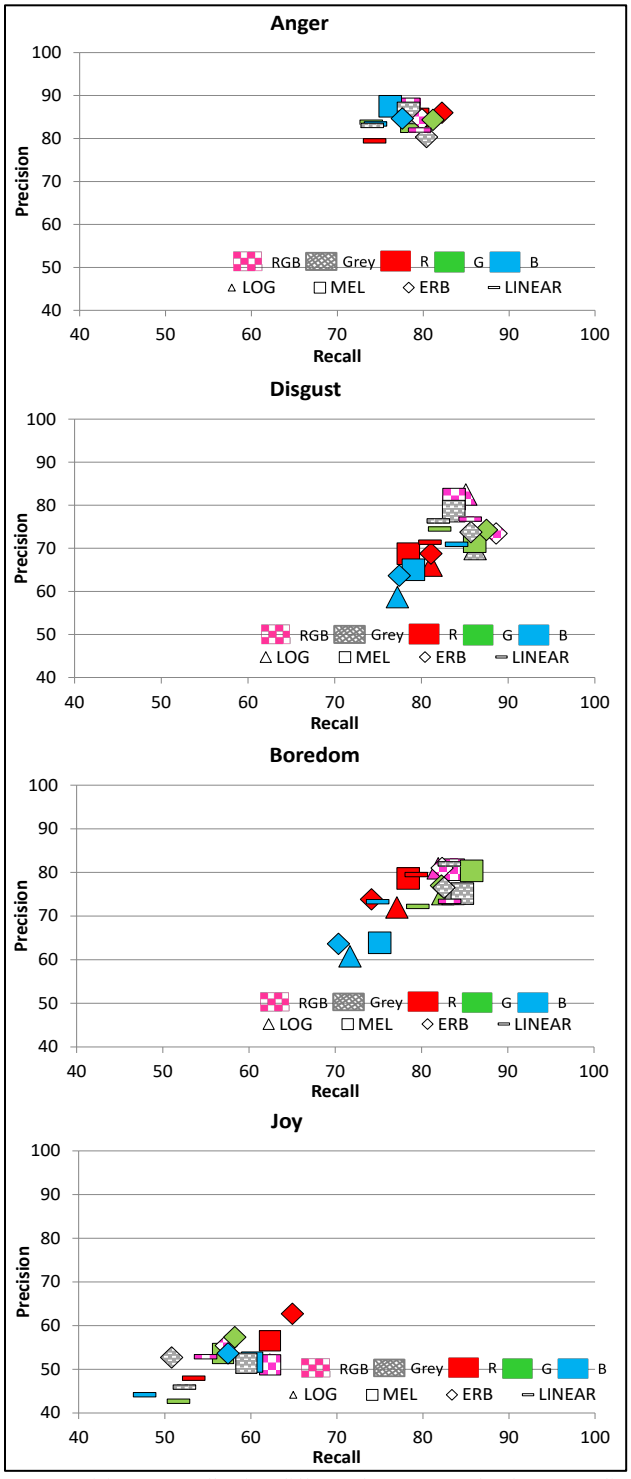


Figure 6. Precision vs. recall using different frequency scales for anger, disgust, joy and boredom. The markers shapes indicate the frequency scale, while their colors indicate types of spectrogram images.

4.4. Distribution of Categorical Emotions on the Amplitude-Frequency Plane

Given the performance analysis of the separate R, G and B images across emotions and frequency scales summarized in Table 3, Figure 8 was made to show the relative positions of seven emotional states of the EMO-DB database on the amplitude-frequency plane. The graph predicts high possibility of confusion in the differentiation between joy and anger, which is consistent with [16,34], as well as, between sadness and fear. Much higher precision/recall scores for anger in Figure 6 suggested that joy was frequently mistaken for anger. Similarly, higher precision/recall scores for sadness in Figure 7 suggested that fear was frequently mistaken for sadness. As expected, neutral speech took central position in relation to other emotions. Both joy and anger were acted in the EMO-DB data by rising the voice level, which is consistent with [10]. Disgust on the other hand, was achieved by variation of the fundamental frequency (F0) and the first formants. Sadness and fear were placed across all amplitudes, but at relatively high frequencies, which could be due to a large number of unvoiced consonants that can be heard when listening to the recordings.

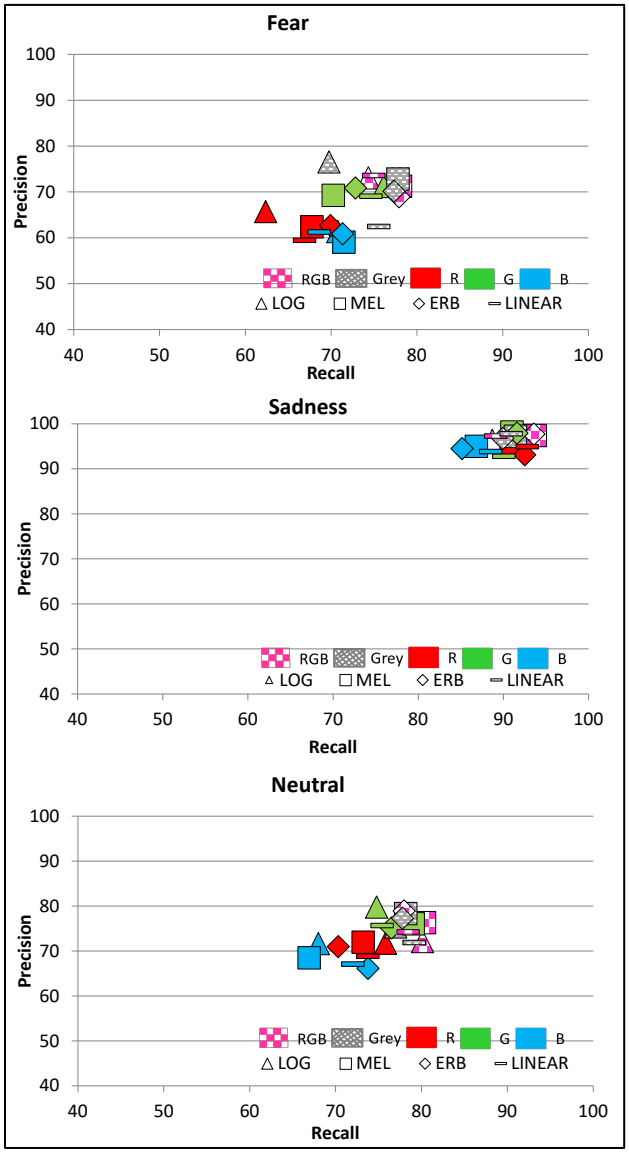


Figure 7. Precision vs. recall using different frequency scales for fear, sadness, and neutral emotional state. The markers shapes indicate the frequency scale, while their colors indicate types of spectrogram images.

Table 3: Amplitude-frequency characteristics of categorical emotions. The total frequency range was 8 kHz, and the total amplitude range was 108 dB.

Emotion	Best Freq. Scale	Best Freq Range	Best CNN Input (3 channels)			Important Amplit. Range
			Ch1	Ch2	Ch3	
Anger	ERB	Medium-Low	R	R	R	High
Disgust	LOG	Low (0.02-2kHz)	R	G	B	All
Joy	ERB	Medium-Low (2-4kHz)	R	R	R	High
Boredom	MEL	Medium-High (4-6kHz)	G	G	G	Medium
Fear	MEL	Medium-High (4-6kHz)	R or Grey scale	G or Grey scale	B or Grey scale	All
Sadness	MEL	Medium-High (4-6kHz)	R	G	B	All
Neutral	ERB	Medium-Low (2-4kHz)	R	G	B	All

Table 4: Statistical significance (p) values obtained when comparing average accuracy of SER based on different types of spectrogram images.

p-values for one-way ANOVA with post-hoc Bonferroni correction, confidence interval: 95%, Bonferroni alpha: 0.05					
	RGB	Grey	R	G	B
RGB	-	0.997	**0.009**	0.868	**0.000**
Grey	0.997	-	0.315	1.000	**0.000**
R	**0.009**	0.315	-	0.367	**0.019**
G	0.868	1.000	0.367	-	**0.000**
B	**0.000**	**0.000**	**0.019**	**0.000**	-

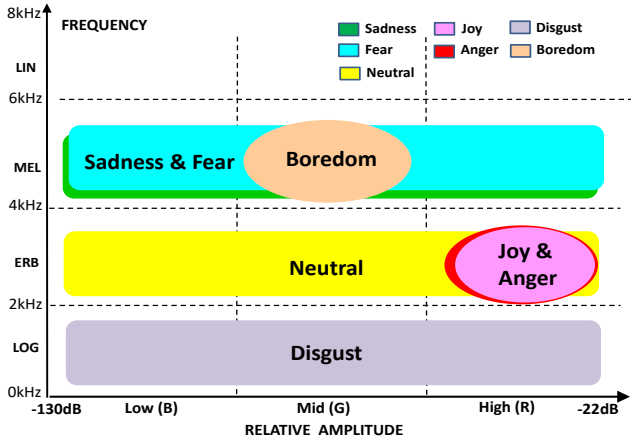


Figure 8. Relative positions of categorical emotions on the amplitude-frequency plane representing acoustic characteristics of emotional speech. These results were obtained using the EMO-DB database

Finally, boredom was localized at mid amplitude levels and high frequencies, which is consistent with very flat, monotonous speech having shallow amplitude modulation. In general, these insights confirmed the expectations, and the fact that they have been generated through machine learning shows that the analysis was valid and consistent with linguistic knowledge. From the perspective of machine learning these insights have direct practical consequences. It is likely that in the future, the SER training process can be enhanced by an adaptive choice of the input features to the network based on optimal configurations given in Table 4.

5. Conclusion

An automatic SER technique based on transfer learning and spectral image classification has been applied to perform an indirect analysis of the amplitude-frequency characteristics of the seven emotional categories represented by the EMO-DB database. Spectrogram images generated on different frequency-scales emphasized different frequency ranges of speech signals, while different color components of the RGB images of spectrograms indicated different values of spectral amplitudes. The analysis provided insights into the amplitude-frequency characteristics of emotional speech. Areas of the amplitude-frequency plane containing cues for different emotions were identified. One of the major limitations of this study is that, the findings apply to acted emotional speech, and only one language (German) was tested. Future works will investigate if the outcomes of the current study are consistent with natural (non-acted) emotional speech, and if they apply across different languages. In addition, factors allowing to efficiently differentiate between highly confused emotions such as for example anger and joy or sadness and fear will be investigated.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was supported by the Defence Science Technology Institute and the Defence Science and Technology Group through its Strategic Research Initiative on Trusted Autonomous Systems, and the Air Force Office of Scientific Research and the Office of Naval Research Global under award number FA2386-17-1-0095.

References

[1] M.N. Stolar, M. Lech, R.S. Bolia, and M. Skinner, "Real Time Speech Emotion Recognition Using RGB Image Classification and Transfer Learning", ICSPCS 2017, 13-16 December 2017, Surfers Paradise, Australia, pp.1-6.

[2] M. Schröder, "Emotional Speech Synthesis: A Review", Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark September 3-7, 2001, pp. 1-4.

[3] F. Eyben, F. Weninger, M. Woellmer, and B. Schuller, "The Munich Versatile and Fast Open-Source Audio Feature Extractor", [Online] Accessed on: Feb 15 2018, Available: https://audeering.com/technology/opensmile/

[4] J.A Bachorovski and M.J. Owren, Vocal expression of emotion: Acoistic properties of speech are associated with emotional intensity and context", Psychological Science, 1995, 6(4), pp. 219-224.

[5] K.R. Scherer, "Non-linguistic indicators of emotion and psychopathology". In: Izard, C.E. (Ed.), Emotions in Personality and Psychopathology. Plenum Press, New York, 1979, pp. 495–529.

[6] K.R. Scherer, 1986. Vocal affect expression: A review and a model for future research. Psychol. Bull. 99 (2), 143–165.

[7] T. Johnstone and K.R. Scherer, 2000. Vocal communication of emotion. In: Lewis, M., Haviland, J. (Eds.), Handbook

[8] K.R. Scherer, "Vocal communication of emotion: A review of research paradigms", Speech Communication 2003, (40), pp. 227–256.

[9] J. Tao, and Y. Kang, "Features importance analysis for emotional speech classification", International Conference on Affective Computing and Intelligent Interaction, ACII 2005: Affective Computing and Intelligent Interaction pp 449-457.
[10] M. Forsell, "Acoustic Correlates of Perceived Emotions in Speech", Master's Thesis in Speech Communication, School of Media Technology, Royal Institute of Technology, Stockholm, Sweden 2007.
[11] Maragos, P., J.F. Kaiser, and T.F. Quatieri, Energy separation in signal modulations with application to speech analysis. Signal Processing, IEEE Transactions on, 1993. 41(10): p. 3024-3051.
[12] D. Ververidis and C. Kotropoulos, Emotional speech recognition: resources, features and methods ", Speech Communication, Volume 48, Issue 9, September 2006, Pages 1162-118.
[13] D. Ververidis, and C. Kotropoulos, Emotional speech recognition: Resources, features, and methods. Speech Communication, 2006. 48(9): p. 1162-1181.
[14] G. J. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. Speech Audio Process., vol. 9, no. 3, pp. 201–216, Mar. 2001.
[15] L. He, M. Lech, N. Maddage, and N. Allen, "Stress detection using speech spectrograms and sigma-pi neuron units", iCBBE, ICNC'09-FSKD'09, 14-16 August 2009 Tianjin, China, pp. 260-264.
[16] L. He, "Stress and emotion recognition in natural speech in the work and family environments", PhD Thesis, RMIT University, Australia, 2010.
[17] A. Nogueiras, J.B. Marino, A. Moreno, and A. Bonafonte,, 2001. Speech emotion recognition using hidden Markov models. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech 2001), Denmark.
[18] F.J. Tolkmitt and K.R. Scherer, 1986. Effect of experimentally induced stress on vocal parameters. J. Exp. Psychol. [Hum. Percept.] 12 (3), 302–313.
[19] R. Banse and K.R. Scherer, 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70 (3), 614–636.
[20] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, 2000. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. 7, 829–837.
[21] T. L. Nwe, S.W. Foo, and L.C. De Silva, 2003. Speech emotion recognition using hidden Markov models. Speech Comm. 41, 603–623.
[22] S.B. Davis and P. Mermelstein (1980), "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." IEEE Transactions on ASSP 28: 357–366.
[23] L. He, M. Lech, and N.B. Allen, "On the Importance of Glottal Flow Spectral Energy for the Recognition", Interspeech 2019, pp. 2346-2349.
[24] B.C.J. Moore and B.R. Glasberg (1983) "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns" The Journal of the Acoustical Society of America (JASA) 1983, 74, pp. 750-753.
[25] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction", 2001 IEEE Signal Processing Magazine, 2001, Volume 18, Issue 1, pp. 32-80.
[26] H. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks", ICSPCS, 14-16 December 2015, Cairns, Australia, pp. 1-6.
[27] H.M. Fayek, M. Lech and L. Cavedon L, "Evaluating deep learning architectures for speech emotion recognition", Neural Networks, March 2017, Special Issue 21, pp. 1-11.
[28] Z. Huang M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN", ACM 2014, November 3-7, 2014, Orlando Florida, USA, pp. 801-804.
[29] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks". In Proceedings of the Signal and Information Processing Association Annual Summit and Conference, Jeju, Korea, December 2016, pp. 1–4.
[30] Q. Mao, M Dong, Z.Huang and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks", IEEE Transactions on Multimedia, 16(8), December 2014, pp. 2203-2213.
[31] S. Prasomphan and K. Mongkut, "Use of neural network classifier for detecting human emotion via speech spectrogram", The 3rd IIAE International Conference on Intelligent Systems and Image Processing 2015, pp. 294-298.
[32] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention", Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, 5-9 March 2017, New Orleans, LA, USA, pp. 2227 - 2231.
[33] A.M. Badshah, J. Ahmad, N. Rahim, and S.W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network", Platform Technology and Service (PlatCon), 2017 International Conference on, 13-15 Feb. 2017, Busan, South Korea, pp. 1-5.
[34] M. Stolar, "Acoustic and conversational speech analysis of depressed adolescents and their parents", 2017, PhD Thesis, RMIT University.
[35] Z. Huang M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN", ACM 2014, November 3-7, 2014, Orlando Florida, USA, pp. 801-804.
[36] Q. Mao, M Dong, Z.Huang and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks", IEEE Transactions on Multimedia, 16(8), December 2014, pp. 2203-2213.
[37] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech", Interspeech 2005, 4-8 September 2005, Lisbon, Portugal pp. 1-4.
[38] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in Neural Information Processing Systems, 2012, pp.1097-1105.
[39] IMAGENET, Stanford Vision Lab, Stanford University, Princeton University 2016 Accessed on: Jan 14 2018 [Online], Available: http://www.image-net.org/
[40] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", Nature, 2015, 521, pp. 436-444.
[41] C.M. Bishop, "Pattern Recognition and Machine Learning", Springer 2006.
[42] MathWorks documentation: Alexnet, Accessed on: Jan 14 2018, [Online], Available: https://au.mathworks.com/help/nnet/ref/alexnet.html?requestedDomain=true
[43] Voicebox, Description of spgrambw Accessed on: Jan 14 2018, [Online], Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/doc/voicebox/spgrambw.html
[44] S.S. Stevens and J. Volkman, "The relation of pitch to frequency: A revised scale" The American Journal of Psychology 1940, 53, pp. 329-353.
[45] B.C.J. Moore and B.R. Glasberg (1983) "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns" The Journal of the Acoustical Society of America (JASA) 1983, 74, pp. 750-753.
[46] H. Traunmüller and A. Eriksson, "The perceptual evaluation of F0-excursions in speech as evidenced in liveliness estimations" The Journal of the Acoustical Society of America (JASA), 1995, 97: pp. 1905 - 1915.
[47] MathWorks; Documentation Jet; Jet colormap array, Accessed on: Jan 14 2018, [Online], Available: https://au.mathworks.com/help/matlab/ref/jet.html?requestedDomain=www.mathworks.com
[48] Matlab Documentation Colormaps: Accessed on: Jan 14 2018, [Online], Available: https://au.mathworks.com/help/matlab/colors-1.html
[49] B. Moore, "An Introduction to the Psychology of Hearing", Sixth Edition, Emerald Group Publishing Limited, 2012.
[50] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge". In Proceedings of Interspeech 2009, pp. 312–315.
[51] B. Schuller, S. Steidl, A. Batliner, F Schiel and J Krajewski, "The Interspeech 2011 speaker state challenge". In Proceedings of Interspeech 2011, August 2011, pp.1-4.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 365-373 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Characteristic, Thermochemical Behaviors and Kinetic of Demineralized and Torrefied Empty Fruit Bunches (EFB)

Nur Nasulhah Kasim[*,1,2], Khudzir Ismail[1,2], Alina Rahayu Mohamed[3], Mohd Azlan Mohd Ishak[1,2], Razi Ahmad[2,4], Wan Izhan Nawawi Wan Ismail[1,2,5]

[1]*Coal and Biomass Energy Research Group, Universiti Teknologi MARA (UiTM), 40450, Shah Alam, Selangor, Malaysia*

[2]*Faculty of Applied Sciences, Universiti Teknologi MARA Malaysia, 40450 Shah Alam, Selangor, Malaysia*

[3]*Faculty of Engineering Technology, Unicity Alam, UniMAP, Padang Besar, 02400 Perlis, Malaysia*

[4]*School of Environmental Engineering, Universiti Malaysia Perlis, 02600 Arau, Perlis, Malaysia*

[5]*Department of chemistry, University York, Heslington, YO10 5DD, United Kingdom*

ARTICLE INFO

Article history:
Received: 13 August, 2018
Accepted: 30 September, 2018
Online: 18 October, 2018

Keywords:
Demineralization
Torrefaction
Empty fruit bunches (EFB)
TGA
Coats-Redfern method

ABSTRACT

A sequential pre-treatment of demineralization and torrefaction, was carried out on palm empty fruit bunches (EFB). EFB and demineralized EFB (DEFB) were torrefied in a vertical tubular reactor in the temperature range of 200 to 280 °C for 30 mins under nitrogen (flow rate:100 mL/min. The pretreated samples were characterized using proximate and ultimate analyses, fuel properties, and Fourier-transform infrared (FTIR) spectroscopy techniques. The thermal and kinetic study on the torrefied samples were carried out using thermogravimetric analysis. The results showed that sequential pre-treatment enhances the properties of solid EFB by increasing the carbon content and reducing the oxygen content with increasing the calorific value. Fuel properties of torrefied samples showed the mass and energy yield decreased, with an increase in energy density. In addition, the FTIR spectra showed the decomposition of hemicellulose occurring for torrefied samples as evidenced by the disappearance of the vibrational features belonging to hydroxyl and carbonyl groups. The kinetic study carried out using Coats-Redfern method on torrefied samples suggested that the activation energy can be transferred by the sequential pre-treatment, indicating that the abundant energy it has can be converted into bio oil of high quality. Apparently, torrefied samples bear high potential to be used as biofuel feedstock when exposed to further thermal decomposition and pyrolysis processes.

1. Introduction

This paper is an extension of work the originally presented in 4th IET Clean Energy and Technology Conference (CEAT 2016) [1]. The extended work was focused on sequential pre-treatment of demineralization and torrefaction of empty fruit bunches (EFB) feedstock. Characteristic and fuel properties of pre-treated EFB were investigated and its kinetic study was performed using Coats-Redfern method.

Depletion of energy resources is a crucial issue which has been challenged worldwide. Fossil fuel energy resources such as petroleum, natural gas and coal are expected to last for approximately 35 years, 37 years and 104 years, respectively [2]. Utilization of agricultural waste seems to assist the shortage of energy resources especially for countries having such resources [3]. Malaysia is the second largest producer of palm oil giving rise to abundant oil palm wastes during the refined, bleached and deodorized (RBD) palm oil processing[4].

Previously, many researchers have focussed on the utilisation of the oil palm waste such as empty fruit bunches (EFB) [5], oil palm fronds (OPF) [6], mesocarp fibre (MF) and palm kernel shell (PKS) [6,7,8] due to their potential benefits for energy sources. However, hygroscopic behavior and higher content of alkali and alkaline earth metals (AAEMs) particularly in EFB limits the

[*]Corresponding Author: Nur Nasulhah Kasim, Universiti Teknologi MARA Malaysia, +604-9882000, nurnasulhah@perlis.uitm.edu.my

utilization of biomass in producing the energy due to vulnerable to biological attack and biodegradation [10]. In addition, nearly 60 % of water contained in wet EFB and the heating value of the dry EFB is roughly half that of palm oil [4].Thus, EFB is unsuitable to be directly burnt for energy production. Therefore, some pre-treatments need to be introduced on EFB prior to pyrolysis to produce high quality of bio oil that contain hydrocarbon content with low carboxylic acid molecules [9,10].

From previous research, feasible pre-treatments on EFB have been investigated such as demineralization and torrefaction in order to enhance the properties of biomass prior to thermal conversion [11,12,13]. Demineralization is an efficient pre-treatment that can be carried out using water, acid or alkaline solution to eliminate the inorganic compounds that assist the secondary reaction during the pyrolysis. Although water reduced about 50 % concentration of inorganic compounds in most of the biomass, chemicals like acid and alkali solution eliminated almost 100 % of the compounds [14,15].

Generally, torrefaction is the often-applied pre-treatment to improve the quality of biomass. The process is carried out in an inert condition at atmospheric pressure between the temperatures of 200-320 °C. This process is a mild version of slow pyrolysis with the aim of reducing the carbonyl, carboxyl, hydroxyl group compounds, and lowering moisture and oxygen contents. In addition, the grindability and hydrophobicity are improved by breaking down the cell wall and fiber structures of the biomass [11]. Washing, drying and size reduction on biomass is a simple pre-treatment that need to be prepared prior to pyrolysis [18]. Basically, an efficient pre-treatment was seen to be easier, economic and less time consuming with simple processing steps. Nevertheless, single pre-treatment such as demineralization or torrefaction alone was as an inappropriate method to obtain high quality bio oil [9,15,17] from raw EFB having high AAEMs and oxygen content.

Previously, application of demineralization and torrefaction sequentially was reported as a novel promising method to reduce the undesirable natural catalyst (AAEMs), moisture, acetyl and acid contents of the biomass samples [10,17]. Therefore, a combination of demineralisation and torrefaction processes was applied to reduce the oxygen, carboxyl, water and ash contents of biomass [19]. Similarly, demineralization using water washing technique combined with torrefaction pre-treatment on microwave pyrolysis of rice husk resulted in large elimination of inorganic compounds which thus, could enhance the quality of fuel characteristic [17].

Pyrolysis of biomass such as EFB, PKS, cotton stalk and rice straw has been performed under dynamic non-isothermal condition in thermogravimetric analyser. To our best knowledge, a few investigations on thermal behaviour and most kinetic studies of biomass pyrolysis have been done using thermogravimetric analyser [4,18,19,20]. In addition, more information or data on kinetic parameters, i.e., activation energy, pre-exponential factor, reaction order and mechanism of reaction can be generated from thermogravimetric analysis [23].

However, there is limited information about physicochemical properties of TEFB and TDEFB at various torrefaction temperatures and very scarce information is available on pyrolysis of demineralized torrefied biomass especially using model free method to calculate activation energy. Thus, the present study was performed to investigate the effect of the sequential application of demineralisation and torrefaction at 200, 220, 240, 260 and 280 °C to enhance the solid fuel qualities.

The basic characteristic of pretreated oil palm EFB including proximate and ultimate analyses, fuel properties and FTIR analysis were elucidated for further utilization of this biomass feedstock. Furthermore, the thermal behaviour and kinetic study of the selected torrefied EFB (i.e. TEFB 240) and torrefied demineralized EFB (i.e. TDEFB 240) were also searched during pyrolysis using thermogravimetric analysis. These data are of significant importance to understand the thermal decomposition and for further effective upgrading of the pre-treatment process. Besides, the thermochemical evolution profiles help designing suitable reactor and setting the optimum condition parameters for the system.

2. Methodology

2.1. Biomass samples preparation

The empty fruit bunches (EFB) were collected from the United Oil Palm Industries Sdn Bhd located in Nibong Tebal, Pulau Pinang, Malaysia. The biomass was rinsed with tap water to remove impurities and dried in oven at 105 °C for 24 h. It was ground using a hammer mill machine into small particle size and sieved using the sieve shaker to 0.5-1.0 cm. Finally, the dried biomass was stored in air-tight containers for further use.

2.2. Sample labels

Untreated EFB, demineralized EFB, torrefied EFB, torrefied demineralized EFB, EFB torrefied at 240 °C and torrefied demineralized EFB at 240 °C were represented as EFB, DEFB, TEFB240 and TDEFB240, respectively.

2.3. Demineralization of EFB (DEFB)

Demineralization of EFB was carried out to eliminate AAEMs that reduce the ash content in the sample. From our preliminary work, sonication for 10 min of EFB in 1 % nitric acid solution at room temperature gave positive results with minimum ash content of 0.74 % in comparison to 4.21 % in untreated EFB [1]. This demineralized EFB sample was used for torrefaction study here.

100 g of EFB was placed in a flask containing 500 mL of 1% nitric acid and was sonicated for 10 min in ultrasonic bath at room temperature. After acid leaching, the sample was filtered and dried at 105 °C for 24 h to a constant weight. Weight measurements for all samples before and after the pre-treatment were carried out using an electronic balance.

2.4. Torrefaction experiment

Torrefaction experiments were performed in a vertical tube fixed bed reactor shown schematically in Figure 1. Each experiment was conducted with 2.0 g of untreated and demineralized EFB samples in inert atmosphere under nitrogen flow (100 mL/min) from ambient temperature to various torrefaction temperatures of 200, 220, 240, 260 and 280 °C, bu keeping at the selected temperature for 30 min. Nitrogen flow was purged before and after 5 min of the experiment and the sample was removed from the reactor after cooling.

2.5. Characterization of EFB, TEFB and TDEFB samples

For each sample, the moisture and ash content, calorific value and ultimate and proximate analyses were carried out. The

moisture content was measured by drying the samples at 105 °C for 4 h [24]. The total ash content and calorific value were determined by ASTM E-1755-01 [25] and ASTM D 2015 – 96 [26] using LECO bomb calorimeter AC-350, respectively. The ultimate analysis of EFB samples was performed according to ASTM D3176 [27] using CHNS-O Flash 2000 with O content determined by difference. Proximate analysis of EFB, TEFB and TDEFB was carried out using Mettler Toledo thermogravimetric analyser (TGA). 10.0 mg of samples were pyrolyzed from ambient temperature to 900 °C at the heating rate of 20 °C/min. The volatile matter was determined using TG/DTG curves and fixed carbon was calculated by difference.

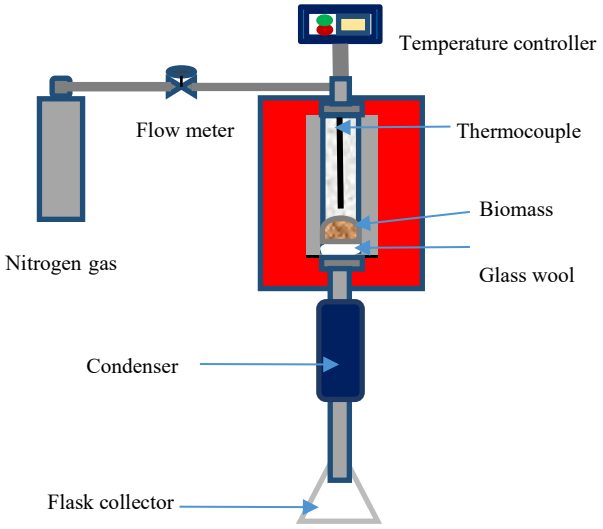


Figure 1. Schematic diagram of torrefaction reactor furnace

The Inductive couple plasma optical emission spectrometer (ICP-OES) was used to determine the inorganic metal content of the EFB and DEFB samples. 300 ± 10 mg of samples were digested with 2.5 mL of 65 % nitric acid and 2.5 mL of 30 % hydrogen peroxide (H_2O_2) in high performance microwave digestion system, Ethos One. The temperature was ramped to 90 °C over 15 min, held for 5 min, then ramped again to 180 °C over 10 min and held for 15 min [11]. The digested samples were analysed using Perkin Elmer Optima 8000 ICP-OES.

Qualitative analysis of functional groups in each EFB samples was conducted using Perkin Elmer Fourier-transform infrared spectroscopy (FTIR) technique. The samples were scanned in the range from 4000-600 cm^{-1} at a resolution of 4 cm^{-1}. For fuel properties analyses, the mass and energy yields of each samples were calculated using equations (1), (2) [28] and energy density equation (3) [29].

$$Mass\ yield\ (Y_{mass}) = \frac{Mass\ after\ torrefaction}{Mass\ of\ raw\ sample} \times 100\% \quad (1)$$

$$Energy\ yield\ (Y_{energy}) = Y_{mass} \times \frac{heating\ value\ of\ torrefied\ sample}{heating\ value\ of\ raw\ sample} \times 100\% \quad (2)$$

$$Energy\ density\ (Y_{energy\ density}) = \frac{Energy\ yield\ (Y_{energy})}{Mass\ yield\ (Y_{mass})} \quad (3)$$

Thermal and kinetic analyses of the EFB, TEFB240 and TDEFB240 samples were carried out using a Mettler Toledo thermogravimetric analyser. 10.0 mg of each EFB sample pyrolysed at the heating rate of 20 °C/min under nitrogen flow (100 mL/min) from ambient temperature to 900 °C.

In this study, Coats-Redfern method which is widely used to follow the thermal decomposition of biomass [4,18,20,28] was applied for modelling the decomposition kinetics of raw EFB, TEFB240 and TDEFB240 samples. By this method, kinetic parameters such as activation energy, E_a and pre-exponential factors, ln A can be determined. In kinetic analysis, the rate of pyrolysis was assumed as:

$$\frac{d\alpha}{dt} = kf(\alpha) \quad (4)$$

In this equation, α represents the progress of reaction or the conversion which occurs between 0.05 to 0.95. α is the progress of reaction during the biomass decomposition which can be defined as:

$$\alpha = \frac{w_0 - w_t}{w_0 - w_f} \quad (5)$$

where w_0, w_t, and w_f are the initial, current (at time t) and final weight of the biomass samples. The rate is constant and dependent on the activation energy (E_a).

$$k = A\exp(-\frac{E_a}{RT}) \quad (6)$$

Considering the assumption of the Coats-Redfern method (2RT/E <<1), and following the first order reaction kinetic, equation (7) represents the kinetic equation used to calculate the activation energy (E_a) and pre-exponential factors, ln A.

$$\ln\left[-\frac{\ln(1-\alpha)}{T^2}\right] = \ln\left[\frac{AR}{\beta E_a}\left[1-\frac{2RT}{E_a}\right]\right] - \frac{E}{RT} \quad (7)$$

According to kinetic equation (7), the plot of $\ln[-\ln(1-\alpha)/T^2]$ against 1/T give a correlation straight line with the first order kinetics. The activation energy and the pre-exponential factor, ln A can be determined from the slope and the intercept of the straight line, respectively [31].

3. Result and Discussion

3.1. Proximate analysis of EFB, DEFB, TEFB and TDEFB

The proximate analysis of EFB, DEFB, TEFB and TDEFB is shown in Table 1. Apparently, the ash content of DEFB dramatically decreased to 0.74 wt. % in comparison with EFB at 4.21 %.

Moreover, the effect of torrefaction on DEFB (represented as TDEFB) can be seen from the increase in the ash content changing in the range from 1.31 – 2.69 %. Likewise, for TEFB, as the torrefaction temperature increased from 200 to 240 °C, the ash content increased slightly and almost doubled at a higher torrefaction temperature of 260-280 °C in comparison with EFB. It was due to the induces partial removal of volatiles from biomass

after the torrefaction pre-treatment. Indeed, torrefaction has a marked effect on the ash content of both the TEFB and TDEFB samples. This implies that sequential application of demineralization and torrefaction on EFB reduced the ash content and offset the negative impact on the increased of ash content due to torrefaction [19].

Table 1. Proximate analysis of EFB, TDEFB and TDEFB

Sample	Proximate analysis (% db)		
	Volatile matter (VM)	Fixed carbon (FC)	Ash
EFB	53.85	41.93	4.21
DEFB	73.60	25.65	0.74
TEFB 200	72.95	22.49	4.54
TEFB 220	72.54	22.64	4.81
TEFB 240	45.95	49.12	4.91
TEFB 260	46.83	44.67	8.49
TEFB 280	46.81	43.59	9.58
TDEFB 200	59.23	39.45	1.31
TDEFB 220	56.27	42.35	1.37
TDEFB 240	63.04	35.08	1.87
TDEFB 260	69.09	29.13	1.76
TDEFB 280	47.98	49.32	2.69

With reference to volatile matter content, TEFB at low torrefaction temperature range of 200 and 220 °C showed a drastic increase by about 35 % in comparison to that of EFB. On contrary, fixed carbon at low torrefaction temperatures (200-220 °C) of TEFB decreased by 46 % which is significant. This trend can be explained by the dehydration and decarboxylation of hemicellulose which occurring at lower torrefaction temperature [30,31]. During the dehydration and decarboxylation, hemicellulose decomposes into H_2O, CO_2, CO and solid char as well as low molecular weight hydrocarbon [28].

Consequently, at the higher torrefaction temperature range from 240 to 280 °C, volatile matter of TEFB was found to decrease with the increase in fixed carbon. This observation is attributed to the initial thermal decomposition of cellulose [34] which completely destroys the cell structure of EFB and causes a further increase in the amount of fixed carbon during high torrefaction temperature [26,33].

As for TDEFB, the volatile matter content became slightly higher with the reduction in fixed carbon content in the torrefaction temperature range from 200 to 260 °C in comparison with EFB. These results indicated a slow rate decomposition of hemicellulose in DEFB during torrefaction. Similar observation was also reported in other studies investigated the influence of inorganic metals on the decomposition of biomass [36]. In contrast, at the torrefaction temperature of 280 °C of TDEFB, the reduction of volatile matter and the increase in fixed carbon content were observed.

This observation might be due to the extensive decomposition of hemicellulose into volatiles and solid products that took place at more severe torrefaction condition [37]. These results indicated that demineralization pre-treatment influenced the degradation of lignocellulosic components in EFB and facilitated the decomposition of hemicellulose and cellulose at different temperatures during torrefaction pre-treatment.

3.2. Ultimate analysis and calorific value of EFB, DEFB, TEFB and TDEFB

The H/C ratio, O/C ratio and calorific value for EFB, DEFB, TEFB and TDEFB are listed in Table 2. The ultimate analysis revealed the alterations of chemical composition of EFB which occurred after the pre-treatment of demineralization and torrefaction. The carbon content was found to increase for single sequential pre-treatment of demineralization and torrefaction. When the torrefaction temperature was increased, the carbon content gradually increased whereas hydrogen and oxygen contents of TEFB and TDEFB samples decreased. These findings can be explained by the effects of dehydration, decarboxylation, and demethanation reactions of hemicellulose [38] and partial depolymerization of cellulose during torrefaction [34].

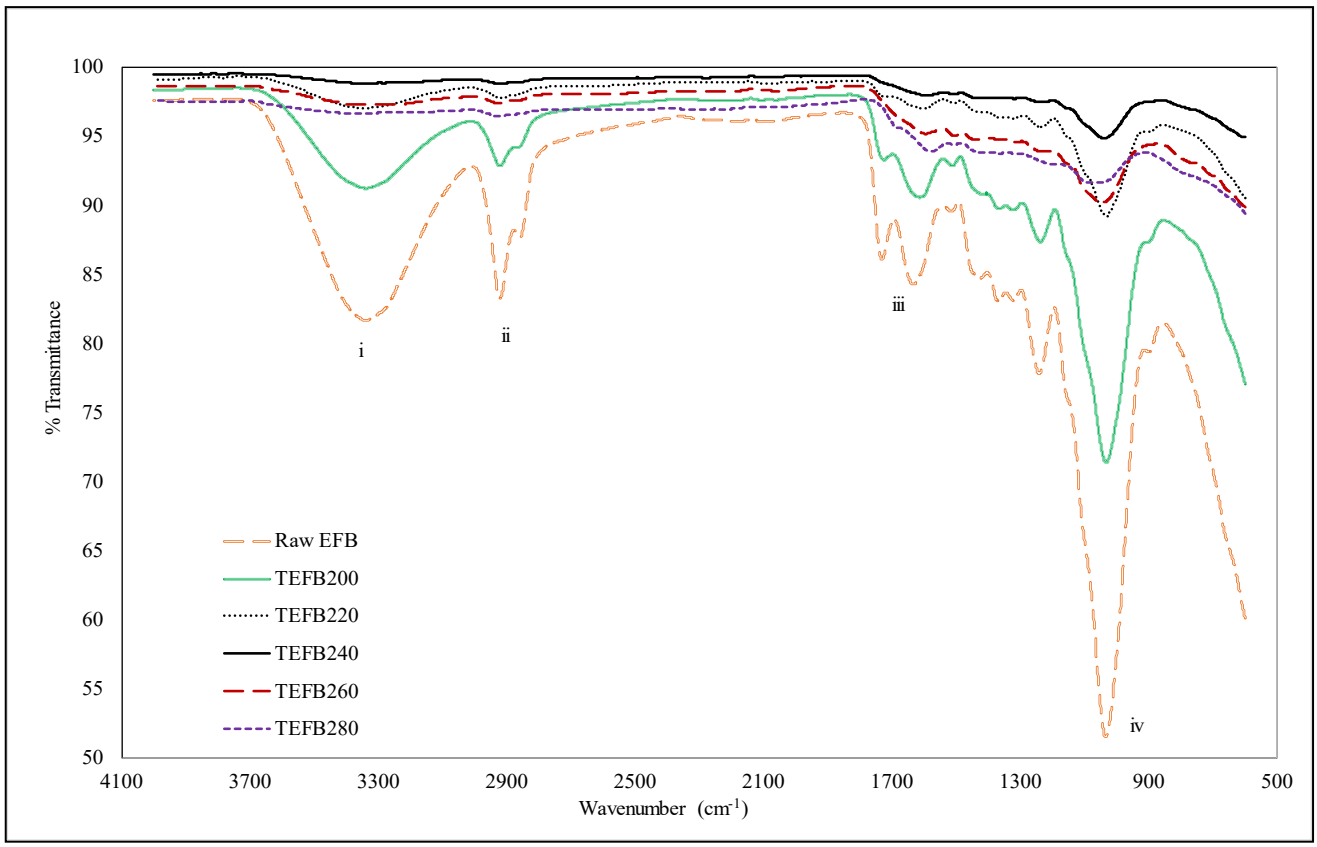
In the literature [35], the atomic ratio of O/C and H/C in biomass are found between 0.4 to 0.8 and 1.2 to 2.0, respectively. As shown in Table 2, H/C ratio of DEFB was slightly lower in comparison with EFB which decreased from 1.66 to 1.52. Similar trend was also observed for H/C and O/C ratio of TEFB and TDEFB when the torrefaction temperature was increased from 200 to 280 °C. This clearly indicated that, moisture and light volatile compounds which contain high hydrogen and oxygen were eliminated from EFB and DEFB, and consequently the carbon content of the materials increased after torrefaction. In general, sulphur and nitrogen content in biomass sample is much lower than that of coal [35].

Both demineralization and torrefaction pre-treatments could remove sulphur (S) and reduce the nitrogen (N) content of EFB (Table 2). This result was supported by previous studies reporting that sulphur and nitrogen content in biomass were reduced after torrefaction [39] and wet torrefaction [37,39].

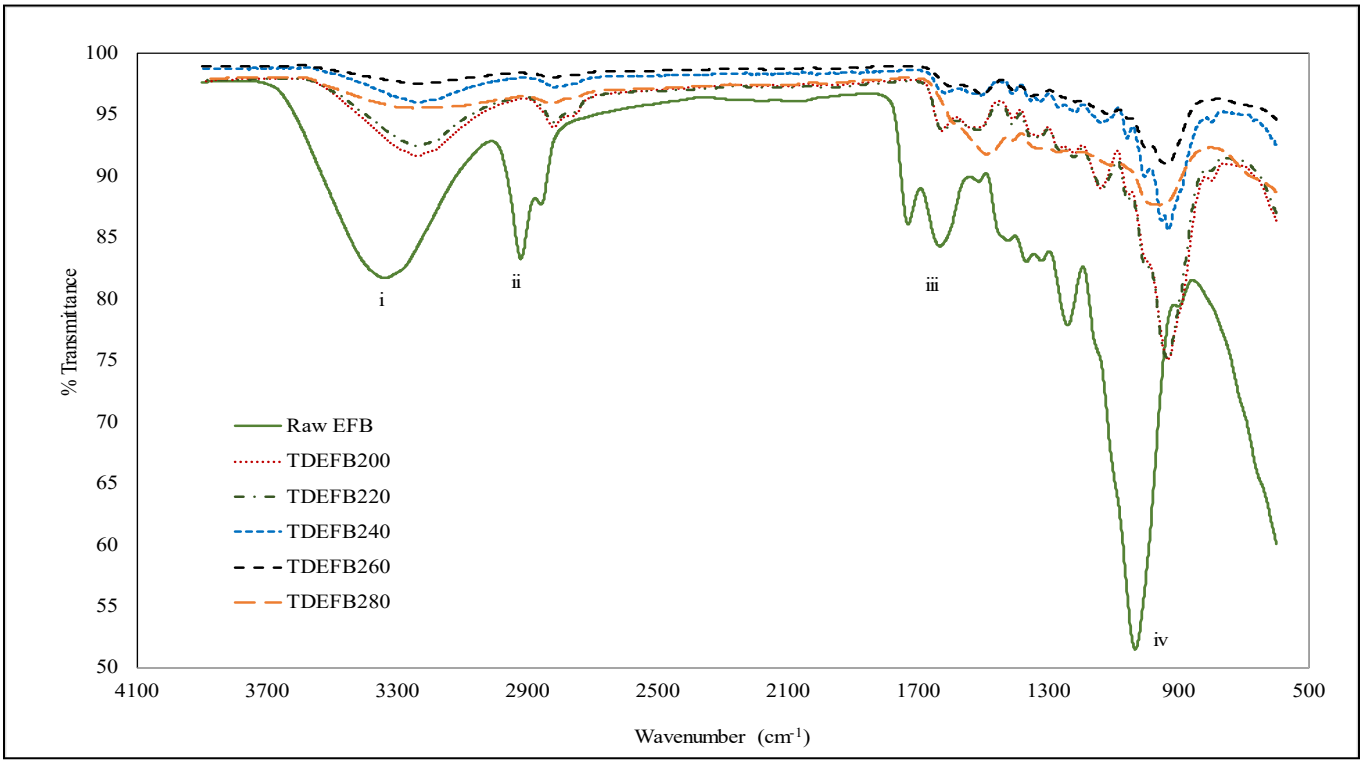
Table 2. Ultimate analysis, atomic ratio and calorific value of raw and pre-treated empty fruit bunch (EFB)

Sample	Ultimate analysis				Calorific value
	O/C	H/C	N/C	S/C	(MJ kg^{-1})
EFB	0.79	1.66	0.03	-	17.80
DEFB	1.52	0.78	0.01		17.30
TEFB 200	0.78	1.53	0.00	-	18.50
TEFB 220	0.74	1.48	0.01		18.90
TEFB 240	0.69	1.27	0.00		20.60
TEFB 260	0.57	1.05	0.00		21.10
TEFB 280	0.48	0.99	0.00		23.50
TDEFB 200	0.73	1.50	0.00	-	17.90
TDEFB 220	0.70	1.43	0.00		18.20
TDEFB 240	0.68	1.39	0.00		19.70
TDEFB 260	0.63	1.27	0.00		21.10
TDEFB 280	0.45	0.99	0.01		23.50

Figure 2 compares the FTIR spectra of EFB with TEFB and TDEFB at various torrefaction temperatures. The FTIR spectra showed four prominent absorption peaks at 3345 cm^{-1} (i), 2920 cm^{-1} and 2850 cm^{-1} (ii), 1650 cm^{-1} and 1710 cm^{-1} (iii) and 1033 cm^{-1} (iv) which are assigned to the vibration of hydroxyl group (-OH), aliphatic methylene group, carbonyl groups (C=O) from aldehydes



(a)



(b)

Figure 2. FTIR spectra of a) EFB and TEFB and, b) EFB and TDEFB at various torrefaction temperatures

and acids, and ethers, phenol and alcohol (C−O, C=C and C−C−O stretching), respectively. The spectra of EFB shows clearly a broad envelope of 3345 cm^{-1} which corresponds to the hydroxyl group. However, this peak was remarkably weakened as the torrefaction temperature was increased due to the removal of moisture through dehydration process.

The peaks at 2920 cm-1 and 2850 cm-1 corresponding to the hydrocarbons and carbonyl species, respectively show similar trend that represents the evolution of carbon dioxide and methane, and extent of the decomposition of hemicellulose for torrefied EFB. Apparently, the intensities of spectral peaks (iii) appearing at 1650 cm-1 and 1710 cm-1 which were assigned to alkene (C=C) and carbonyl (C=O), respectively decreased for both TEFB and TDEFB samples in comparison to EFB. This observation indicates that oxygen and hydrogen have been removed through decarboxylation and dehydration of hemicellulose during torrefaction and demineralisation of EFB. This is in agreement with similar findings reported by other researchers [39, 40].

Previous research [41] reported that torrefaction pre-treatment could reduce the acid content of bio oil due to the removal of acetyl group in hemicellulose which is the primary source of acetic acid. Torrefaction of EFB and DEFB samples above 240 °C decreased the intensity of the peak originating form acetyl group (1710 cm-1), which is observation attributed to deacetylation of hemicellulose. Therefore, torrefaction temperature of 240 °C were selected for TEFB and TDEFB in bio fuel feedstock. This is due to the fact that at this temperature, the percentages of torrefaction yield of TEFB240 and TDEFB240 were found higher than that of EFB and DEFB samples at the torrefaction temperatures of 260 and 280 °C with less acetyl group content

The Van Krevelen diagram can be used to analyse the characteristics of torrefied biomass based on the relative increase in carbon content while decreasing the relative amounts of hydrogen and oxygen [35].

With reference to Figure 3a, the Van Krevelen diagram of TEFB and EFB shows the reduction in elementary composition of oxygen and hydrogen with the increase of torrefaction temperature. Comparable tendency was also observed for TDEFB as shown in Figure 3b. The changes in elementary composition of both torrefied samples resemble sub-bituminous coal referring to the Van Krevelen plot. These results indicate that some alterations in fuel properties [28] that enhance the quality of torrefied EFB as a solid fuel feedstock [34].

3.3. The effect of sequential pre-treatment on fuel properties of EFB, TEFB and TDEFB

The mass and energy yields and energy density were calculated using equations 1, 2 and 3, respectively and the data are illustrated in Figure 4. Indeed, it is seen from Figure 4 that torrefaction pre-treatment had an important impact on the mass and energy yields and energy density. With the increase in torrefaction temperature from 200 to 280 °C, the mass and energy yields for TDEFB declined gradually from 77.88 to 34.96 % and 80.94 to 46.15 %, respectively. The same tendency of reduction was also observed for the mass and energy yields of the TDEFB sample in the range from 86.01 to 45.89 % and from 86.49 % to 60.59 %, respectively.

However, mass and energy yields of TDEFB were slightly higher than that of TEFB. These results may be associated with the trending of volatile matter of proximate analysis as shown in Table 1 in previous section. TEFB suffered from obvious decreases of mass and energy yields specifically at the temperatures above 240 °C. Meanwhile, TDEFB was able to retain a drastic loss of mass and energy yields prior to the pre-treatment. In addition, the energy density of TEFB and TDEFB increased in the range from 1.01 to 1.34 with the increased of the torrefaction temperature from 200 to 280 °C. This was due to the major decomposition of hemicellulose and partial decomposition of cellulose during torrefaction resulting in an increase in calorific value, and is similar to the results reported by other researchers [41,39,32].

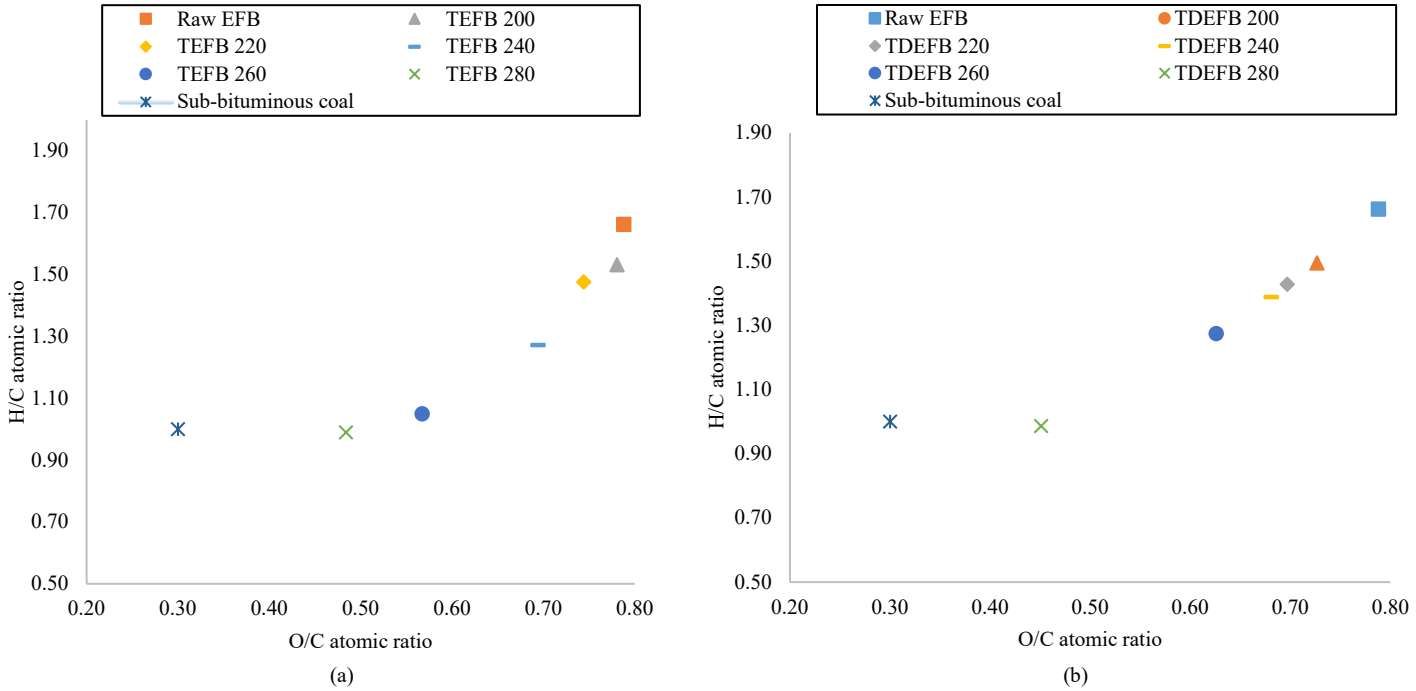
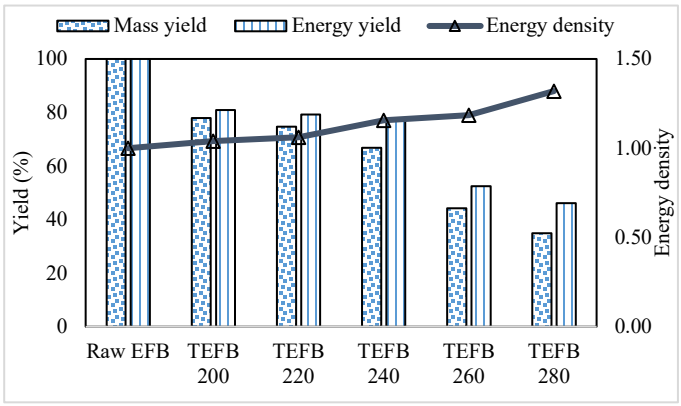
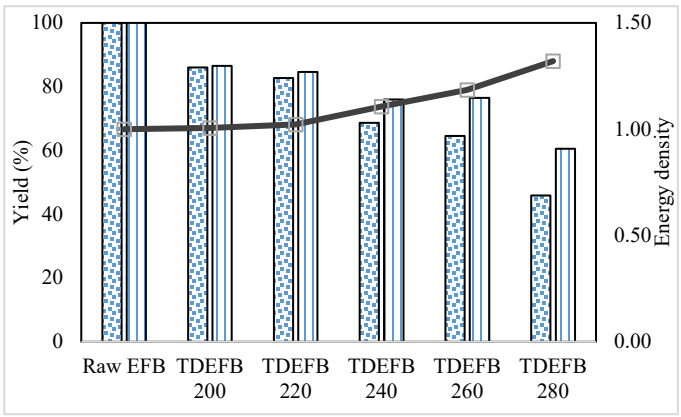


Figure 3. Van Krevelen diagram for chemical composition of a) EFB and TEFB at various torrefaction temperatures and b) EFB and TDEFB at various torrefaction temperatures (sub-bituminous coal as reference)



(a)



(b)

Figure 4. Mass and energy yield and calorific value a) TEFB and b) TDEFB

According to our previous work, pre-treated palm wastes such as EFB and palm kernel shell (PKS) using physical and chemical pre-treatments were identified the appropriate solid biofuel for further thermal conversions such as combustion, pyrolysis, gasification and other applications. Further, it also promote to the enrichment of energy security, decrease in CO2 emissions, sustenance to the development of renewable energy and high advantageous to the agricultural and forestry activity [44,45].

3.4. Thermogravimetric analysis

Based on the results obtained from proximate and ultimate analyses, fuel properties and FTIR data of EFB, TEFB and TDEFB, both TEFB 240 and TDEFB 240 were obtained following the pyrolysis process. Thermogravimetric analyser (TGA) was used to observe the thermal behaviour and to obtain the kinetic parameters of all samples. The weight loss (TG) and the differential weight loss (DTG) curves of EFB, TEFB 240 and TDEFB 240 are illustrated in Figure 5 and 6, respectively.

From Figure 5, the first weight loss is identified at approximately 128 °C for all samples. This corresponds to the removal of water molecules which are adsorbed on the surface of EFB and is shown by the first small peak in Figure 6. The second peak shows that the major weight loss occurs from 200 to 350 °C for EFB and from 290 to 350 °C for TEFB 240 and TDEFB 240. This could be due to the active pyrolysis of hemicellulose and cellulose portion in EFB. The cellulose and hemicelluloses contents of EFB are higher than those of TEFB 240 and TDEFB 240. Therefore, the active pyrolysis process began at a much lower temperature for EFB compared to TEFB 240 and TDEFB 240. This devolatilization stage of EFB resulted in changes via occurrence of simultaneous reactions such as depolymerization, decarboxylation, dehydration and decarbonylation [6, 42].

The TEFB 240 sample had undergone single pre-treatment of torrefaction at 240 °C whereas the TDEFB 240 sample had experienced sequential demineralization and torrefaction at 240 °C during pre-treatment stage which resulted in the cleavage of C-O bonds in hemicelluloses and decrease in cellulosic content of EFB. Therefore, the active pyrolysis processes of TEFB 240 and TDEFB 240 are contributed by the remaining lignocellulosic content of EFB that shifts the process to a maximum region. It can be seen from Figure 6 that, the peak temperature shifts from 318 to 336 °C first and finally to 354 °C for EFB, TEFB 240 and TDEFB 240, respectively. This trend is similar to previous study [20], where the main weight loss was observed in the range from 250 to 450 °C with the corresponding DTG curve appearing from 350 to 390 °C and this showed the decomposition temperature regions represented by hemicellulose, cellulose and lignin.

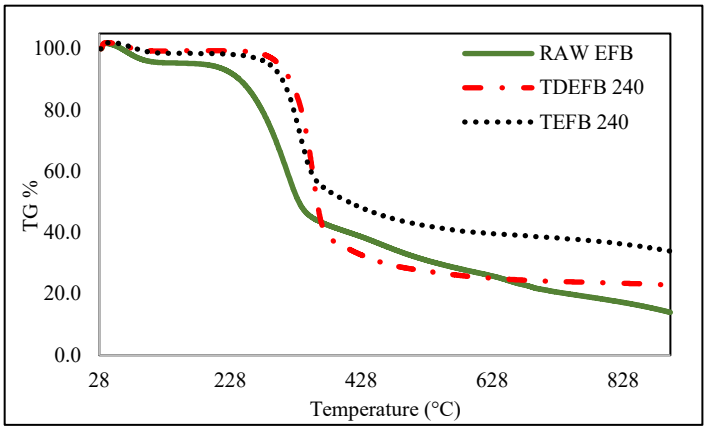


Figure 5. The weight loss (TG) of EFB, TEFB and TDEFB at heating rate of 20 °C min-1 under inert condition

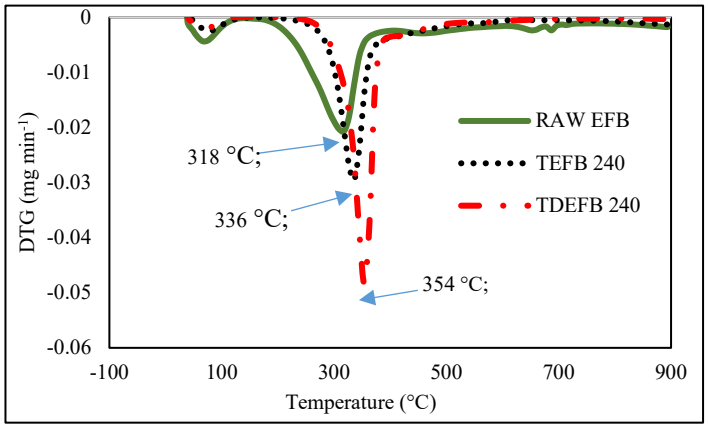


Figure 6. The weight loss rate (DTG) of EFB, TEFB and TDEFB at a heating rate of 20 °C min^{-1} under inert condition

The third stage is observed at a temperature range from 350 to 600 °C in Figure 5. This is a slow weight loss taking place over a wide range of temperature and is due to the composition of lignin in the samples. In summary, three temperature regions are clearly seen for decomposition of lignocellulosic of raw EFB, TEFB 240 and TDEFB 240

Table 3. The kinetic parameters for raw EFB, TEFB 240 and TDEFB 240 pyrolysis obtained by Coats-Redfern model

Samples	Equation	Activation energy, E_a (kJ mol^{-1})	R^2	Pre-exponential factor, ln A (min^{-1})
EFB	y = -8588x +1.8068	71.40	0.9956	10.87
TEFB 240	y = -14573x + 10.81	121.16	0.9898	20.40
TDEFB 240	y = -16592x + 13.419	137.42	0.9888	23.13

3.5. Thermal degradation kinetic analysis

The thermogravimetric data obtained at a heating rate of 20°C min^{-1} were used to calculate the kinetics parameters such as activation energy and pre-exponential factor, ln A for EFB, TEFB 240 and TDEFB 240 using the Coats-Redfern method of the data obtained in the temperature range from 200 to 400 °C. The plot of ln $(-\ln(1-\alpha)/T^2)$ vs. 1/T is illustrated in Figure. 7. The corresponding values of activation energy (E_a), correlation coefficient (R^2) and intercepts are presented in Table 3.

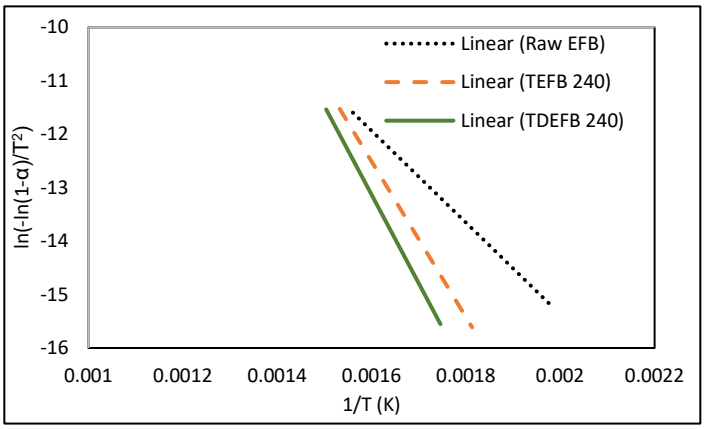


Figure 7. The plot of regression line was produced from value of ln[-ln(1-α)/T2 against 1/T for EFB, TEFB 240 and TDEFB 240

The values of R^2 are above 0.96 indicating the present of a strong correlation between dependable and undependable parameters. From Table 3, the activation energy values for pyrolysis of TDEFB 240 and TEFB 240 were determined as 137.42 and 121.16 kJ mol^{-1}, respectively. These values are higher than that for pyrolysis of EFB which is 71.40 kJ mol^{-1}. The low activation energy of EFB pyrolysis is attributed to the devolatilization of hemicellulosic component [47].

The higher values of activation energy for pyrolysis of torrefied samples may be explained by the fact that pyrolysis of torrefied samples containing mainly C-C networks occurs at higher temperature which can be accompanied by a higher activation energy [20]. During torrefaction of TEFB 240 and TDEFB 240, the hemicellulosic component decomposes and results in the cleavage of C-O and C-H bonds. The removal of oxygen atoms occurs also during this stage either in the form of water or carbon monoxide.

Conclusion

In this work, the effects of sequential pre-treatment consist of demineralization and torrefaction on oil palm empty fruit bunch have been studied. This pre-treatment effectively enhances the quality of bio fuel feedstock by increasing the carbon content, eliminating the AAEMs content and reducing oxygen content. It was found that, this sequential pre-treatment improves the energy density of the selected torrefied samples by decreasing oxygen content which contributed in increasing the calorific value. This results also can also be supported by decreasing the intensity of hydroxy and carbonyl peaks that was observed in FTIR spectra of the samples. According to thermogravimetric analysis, both samples showed the increasing in C-C networks that represent the removal of hemicellulose and rich in cellulose content. These results can be supported by increasing the activation energy of TEFB 240 and TDEFB 240 using Coats-Redfern method of kinetic study. Thus, this sequential pre-treatment is suggested to be the most effective approach for upgrading the quality of solid fuel feedstock for further thermal conversion process for the future.

Acknowledgment

The authors acknowledge Universiti Teknologi MARA and Universiti Malaysia Perlis (UniMAP) for providing laboratory and instrumentation facilities during the research work. The authors would like to thank also to United Oil Palm Industries Sdn Bhd located in Nibong Tebal, Pulau Pinang, Malaysia for supplying the oil palm EFB sample.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] N. N. Kasim, K. Ismail, M. A. M. Ishak, R. Ahmad, A. R. Mohamed, and W. I. Nawawi, "Demineralization of oil palm empty fruit bunch (EFB) intended as a high quality bio-oil feedstock," 4th IET Clean Energy Technol. Conf. (CEAT 2016), 107 (6.), 2016. https://doi:10.1049/cp.2016.1364.

[2] B. R. Singh and O. Singh, "Global Trends of Fossil Fuel Reserves and Climate Change in the 21st Century," Foss. Fuel Environ., 167–192, 2012. https://doi:10.5772/38655.

[3] M. V. Kok and E. Özgür, "Thermal analysis and kinetics of biomass samples," Fuel Process. Technol.,106, 739–743, 2013. https://doi:10.1016/j.fuproc.2012.10.010.

[4] N. Abdullah and F. Sulaiman, "The oil palm wastes in Malaysia," in Biomass Now – Sustainable Growth and Use, 75–100, 2013. https://doi:10.5772/55302.

[5] A. R. Mohamed and Z. Hamzah, "An alternative approach for the screening of catalytic empty fruit bunch (EFB) pyrolysis using the values of activation energy from a thermogravimetric study," React. Kinet. Mech. Catal.,114, 529–545, 2014. https://doi:10.1007/s11144-014-0798-8.

[6] X. Y. Lim and J. M. Andrésen, "Pyro-catalytic deoxgenated bio-oil from palm oil empty fruit bunch and fronds with boric oxide in a fixed-bed reactor," Fuel Process. Technol., 92(9), 1796–1804, 2011. https://doi:10.1016/j.fuproc.2011.04.033.

[7] M. a. Aziz, K. M. Sabil, Y. Uemura, and L. Ismail, "A study on torrefaction of oil palm biomass," Journal of Applied Sciences, 12(11). 1130–1135, 2012. https://doi:10.3923/jas.2012.1130.1135.

[8] J. Mabrouki, M. A. Abbassi, K. Guedri, A. Omri, and M. Jeguirim, "Simulation of biofuel production via fast pyrolysis of palm oil residues," Fuel,159,819–827, 2015. https://doi:10.1016/j.fuel.2015.07.043.

[9] Y. Uemura, W. N. Omar, T. Tsutsui, and S. B. Yusup, "Torrefaction of oil palm wastes," Fuel, 90(8), 2585–2591, 2011. https://doi:10.1016/j.fuel.2011.03.021.

[10] S. Matali, N. A. Rahman, S. S. Idris, N. Yaacob, and A. B. Alias, "Lignocellulosic Biomass Solid Fuel Properties Enhancement via Torrefaction," Procedia Eng., 148, 671–678, 2016.

https://doi:10.1016/j.proeng.2016.06.550.
[11] T. Wigley, A. C. K. Yip, and S. Pang, "The use of demineralisation and torrefaction to improve the properties of biomass intended as a feedstock for fast pyrolysis," J. Anal. Appl. Pyrolysis, 113, 296–306, 2015. https://doi:10.1016/j.jaap.2015.02.007.
[12] T. Wigley, A. C. K. Yip, and S. Pang, "Pretreating biomass via demineralisation and torrefaction to improve the quality of crude pyrolysis oil," Energy, 109, 481–494, 2016. https://doi:10.1016/j.energy.2016.04.096.
[13] N. Abdullah, F. Sulaiman, and H. Gerhauser, "Characterisation of oil palm empty fruit bunches for fuel application," J. Phys. Sci., 22(1), 1–24, 2011.
[14] K. Kaminaka, Y. Matsumura, W. Noaman Omar, and Y. Uemura, "Process evaluation for torrefaction of empty fruit bunch in malaysia," J. Japan Pet. Inst., 57(2), 88–93, 2014.
[15] Y. Uemura, W. Omar, N. A. Othman, S. Yusup, and T. Tsutsui, "Torrefaction of oil palm EFB in the presence of oxygen," Fuel, 103, 156–160, 2013. https://doi:10.1016/j.fuel.2011.11.018.
[16] A. Shariff, N. S. M. Aziz, and N. Abdullah, "Slow Pyrolysis of Oil Palm Empty Fruit Bunches for Biochar Production and Characterisation," J. Phys. Sci., 25(2), 97–112, 2014.
[17] S. Zhang, Q. Dong, L. Zhang, Y. Xiong, X. Liu, and S. Zhu, "Effects of water washing and torrefaction pretreatments on rice husk pyrolysis by microwave heating.," Bioresour. Technol., 193, 442–448, 2015. https://doi:10.1016/j.biortech.2015.06.142.
[18] J. M. Reckamp, R. A. Garrido, and J. A. Satrio, "Selective pyrolysis of paper mill sludge by using pretreatment processes to enhance the quality of bio-oil and biochar products," Biomass and Bioenergy, 71, 235–244, 2014. https://doi:10.1016/j.biombioe.2014.10.003
[19] D. Chen, J. Mei, H. Li, Y. Li, M. Lu, T. Ma, and Z. Ma, "Combined pretreatment with torrefaction and washing using torrefaction liquid products to yield upgraded biomass and pyrolysis products," Bioresour. Technol., 228, 62–68, 2017. https://doi:10.1016/j.biortech.2016.12.088.
[20] L. Cao, X. Yuan, L. Jiang, C. Li, Z. Xiao, Z. Huang, X. Chen, G. Zeng, and H. Li, "Thermogravimetric characteristics and kinetics analysis of oil cake and torrefied biomass blends," Fuel, 175, 129–136, 2016. https://doi:10.1016/j.fuel.2016.01.089.
[21] M. Kopczyński, A. Plis, and J. Zuwała, "Thermogravimetric and kinetic analysis of raw and torrefied biomass combustion," Chem. Process Eng. - Inz. Chem. i Proces., 36(2), 209–223, 2015. https://doi:10.1515/cpe-2015-0014.
[22] P. Ninduangdee, V. I. Kuprianov, E. Y. Cha, R. Kaewrath, P. Youngyuen, and W. Atthawethworawuth, Thermogravimetric Studies of Oil Palm Empty Fruit Bunch and Palm Kernel Shell: TG/DTG Analysis and Modeling, 79, 2015. https://doi:10.1016/j.egypro.2015.11.518.
[23] G. Wang, J. Zhang, J. Shao, Z. Liu, G. Zhang, T. Xu, J. Guo, H. Wang, R. Xu, and H. Lin, "Thermal behavior and kinetic analysis of co-combustion of waste biomass/low rank coal blends," Energy Convers. Manag., 124, 414–426, 2016. https://doi:10.1016/j.enconman.2016.07.045.
[24] S. D. Stefanidis, E. Heracleous, D. T. Patiaka, K. G. Kalogiannis, C. M. Michailof, and A. A. Lappas, "Optimization of bio-oil yields by demineralization of low quality biomass," Biomass and Bioenergy, 83, 105–115, 2015. https://doi:10.1016/j.biombioe.2015.09.004.
[25] P. Basu and P. Basu, "Chapter 13 – Analytical Techniques," in Biomass Gasification, Pyrolysis and Torrefaction, 439–455, 2013. https://doi:10.1016/B978-0-12-396488-5.00013-7.
[26] S. S. Idris, N. A. Rahman, K. Ismail, A. B. Alias, Z. A. Rashid, and M. J. Aris, "Investigation on thermochemical behaviour of low rank Malaysian coal, oil palm biomass and their blends during pyrolysis via thermogravimetric analysis (TGA)," Bioresour. Technol., 101(12), 4584–4592, 2010. https://doi:10.1016/j.biortech.2010.01.059.
[27] K. Singh and J. Zondlo, "Characterization of fuel properties for coal and torrefied biomass mixtures," J. Energy Inst., 2016. https://doi:10.1016/j.joei.2016.05.012.
[28] T. G. Bridgeman, J. M. Jones, A. Williams, and D. J. Waldron, "An investigation of the grindability of two torrefied energy crops," Fuel, 89,(12), 3911–3918, 2010. https://doi:10.1016/j.fuel.2010.06.043.
[29] B. Bevan, A. Ahmad, A. Johari, T. Amran, and T. Abdullah, "Torrefaction of Pelletized Oil Palm Empty Fruit Bunches," in The 21st International Symposium on Alcohol Fuels – 21st ISAF, 15–19, 2015.
[30] M. A. Islam, M. Asif, and B. H. Hameed, "Pyrolysis kinetics of raw and hydrothermally carbonized Kar (Pongamia pinnata) fruit hulls via thermogravimetric analysis," Bioresour. Technol., 179, 227–233, 2015. https://doi:10.1016/j.biortech.2014.11.115.
[31] A. R. Mohamed, Z. Hamzah, M. Z. M. Daud, and Z. Zakaria, "The Effects of Holding Time and the Sweeping Nitrogen Gas Flowrates on the Pyrolysis of EFB using a Fixed–Bed Reactor," Procedia Eng., 53, 185–191, 2013. https://doi:10.1016/j.proeng.2013.02.024.
[32] S. Ren, H. Lei, L. Wang, Q. Bu, S. Chen, and J. Wu, "Thermal behaviour and kinetic study for woody biomass torrefaction and torrefied biomass pyrolysis by TGA," Biosyst. Eng., 116(4), 420–426, 2013. https://doi:10.1016/j.biosystemseng.2013.10.003.
[33] J. Meng, J. Park, D. Tilotta, and S. Park, "The effect of torrefaction on the chemistry of fast-pyrolysis bio-oil.," Bioresour. Technol., 111, 439–446, 2012. https://doi:10.1016/j.biortech.2012.01.159.
[34] Y. Yue, H. Singh, B. Singh, and S. Mani, "Torrefaction of Sorghum Biomass to Improve Fuel Properties," Bioresour. Technol., 232, 372–379, 2017. https://doi:10.1016/j.biortech.2017.02.060.
[35] W.-H. Chen, J. Peng, and X. T. Bi, "A state-of-the-art review of biomass torrefaction, densification and applications," Renew. Sustain. Energy Rev., 44, 847–866, 2015. https://doi:10.1016/j.rser.2014.12.039.
[36] A. Saddawi, J. M. Jones, and A. Williams, "Influence of alkali metals on the kinetics of the thermal decomposition of biomass," Fuel Process. Technol., 104, 189–197, 2012. https://doi:10.1016/j.fuproc.2012.05.014.
[37] P. Basu, Torrefaction. Elsevier Inc., 2013. https://doi:10.1016/B978-0-12-396488-5.00004-6.
[38] S. Zhang, T. Chen, W. Li, Q. Dong, and Y. Xiong, "Bioresource Technology Physicochemical properties and combustion behavior of duckweed during wet torrefaction", 218, 1157–1162, 2016.http:// 10.1016/j.enconman.2016.07.045
[39] W.-H. Chen, S.-W. Du, C.-H. Tsai, and Z.-Y. Wang, "Torrefied biomasses in a drop tube furnace to evaluate their utility in blast furnaces," Bioresour. Technol., 111, 433–438, 2012. https://doi:10.1016/j.biortech.2012.01.163.
[40] A. Zheng, Z. Zhao, Z. Huang, K. Zhao, G. Wei, X. Wang, F. He, and H. Li, "Catalytic fast pyrolysis of biomass pretreated by torrefaction with varying severity," Energy & Fuels, 28, 5804–5811, 2014. https://doi:10.1021/ef500892k.
[41] S. Wang, G. Dai, B. Ru, Y. Zhao, X. Wang, J. Zhou, Z. Luo, and K. Cen, "Effects of torrefaction on hemicellulose structural characteristics and pyrolysis behaviors," Bioresour. Technol., 218, 1106–1114, 2016. https://doi:10.1016/j.biortech.2017.02.120.
[42] Y. Joshi, M. Di Marcello, and W. de Jong, "Torrefaction: Mechanistic study of constituent transformations in herbaceous biomass," J. Anal. Appl. Pyrolysis, 115, 353–361, 2015. https://doi:10.1016/j.biortech.2016.07.075.
[43] D. Chen, K. Cen, X. Jing, J. Gao, C. Li, and Z. Ma, "An approach for upgrading biomass and pyrolysis product quality using a combination of aqueous phase bio-oil washing and torrefaction pretreatment," Bioresour. Technol., 233, 150–158, 2017.
[44] N. N. Kasim, K. Ismail, M. A. M. Ishak, R. Ahmad, A. R. Mohamed, and W. I. Nawawi, "Palm deto : resourceful , sustainable and favorable solid biofuel," 2, 391–397, 2017.
[45] R. Ahmad, M. A. M. Ishak, K. Ismail, and N. N. Kasim, " Co- palm: versatile , clean and enriched energy fuel," 2, 327–331, 2017.
[46] S. S. Mohtar, T. N. Z. Tengku Malim Busu, A. M. Md Noor, N. Shaari, and H. Mat, "An ionic liquid treatment and fractionation of cellulose, hemicellulose and lignin from oil palm empty fruit bunch," Carbohydr. Polym., 166, 291–299, 2017. https://doi:10.1016/j.carbpol.2017.02.102.
[47] H. Yang, R. Yan, H. Chen, D. H. Lee, and C. Zheng, "Characteristics of hemicellulose, cellulose and lignin pyrolysis," Fuel,86(12–13), 1781–1788, 2007. https://doi:10.1016/j.fuel.2006.12.013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 4, 372-377 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

IR Sensing Embedded System Development for Prototype Mobile Platform and Multisensory Data Fusion for Autonomous Convoy

Hubert Bryan Riley*, David Solomon Raj Kondru, Mehmet Celenk

School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 45701, USA

A R T I C L E I N F O

Article history:
Received: 22 May, 2018
Accepted: 10 August, 2018
Online: 25 August, 2018

Keywords:
Autonomous vehicles
Self-driving cars
Embedded systems
IR sensing
Sensor fusion

A B S T R A C T

Advanced sensing technologies are providing for greater capabilities to discern and classify details of objects as they appear in actual environments as experienced by nonprofessional drivers. Distinctive geometric configurations of new sensory devices including but not limited to infrared (abbreviated as IR) and LIDAR sensory units are appearing as cost effective data acquisition systems for environment sensing and presenting. Here, we describe a novel IR sensory-based autonomous vehicle guidance and its associated convoy unit. The underlying systems is first presented in a general system model and the experimental test results have been elaborated to demonstrate the usefulness of the presented prototype in futuristic auto industries and its supporting branches, respectively.

1. Introduction

As edge cities continue to emerge, drivers are seeking to increase their productivity during commute time and self-driving vehicles are expected to be facilitating platforms. The Boston Consulting Group and the World Economic Forum have been collaborating for the past several years and forecasting that 14% of residents will ultimately expect safe autonomous vehicles [1]. This paper examines the performance of an infrared (IR) sensing using a prototype mobile platform for autonomous vehicle convoy following. The IR spectrum sensing for autonomous vehicle applications is a compelling approach to detecting objects in the path of vehicle travel. Since cost, reliability, and detection performance are the critical criteria for technology production, the study of performance is the primary interest in this work. Beyond the practical benefits, autonomous cars could contribute $1.3 trillion in annual savings to the US economy alone, with global savings estimated at over $5.6 trillion[2]along with a technology adaptation time line as shown in Figure 1. Design of self-driving or autonomous vehicles with smart sensing systems will positively impact planning decisions for optimal traffic flow, minimize traffic congestions, and alleviate human error leading to personal injuries or property damage. As commuters elevate expectations for their time efficiencies during travels it is expected that technology developers reduce the potential for crashes due to driver fatigue, poor maneuvering decisions, and negligence.[3],[4],[5].

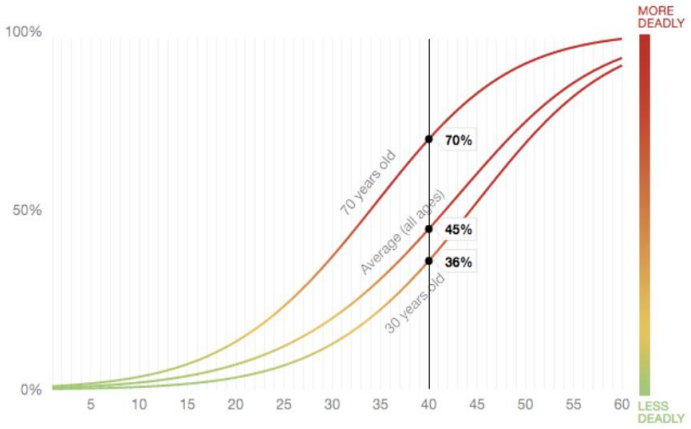


Figure 1. Trends of serious injury vs. vehicle speed in mph.

The curves in Figure 1 is derived comes from typical driving research conducted by AAA Foundation which clearly indicated the probability of getting seriously injured increases with increases in driving speeds. The research also notes these trends depend on several other factors such as size of car, angle of impact, and perhaps age. Autonomous driving systems can aid in reducing or eliminating in the severity of slope in the curves as computed from a small sample size. Cost appears to be a primary driver factor at the onset systems design requirements. To this end, low-cost IR

*Corresponding Author H. Bryan Riley, rileyhbryan@ieee.org

sensing is the motivated attribute to tackle with a vehicle subsystem implementation by noting that IR sensory data acquisition apparatus is capable of sensing surroundings by means of measuring the temperature differences between two vehicles in motion. The vehicle sensor set with field of view (FOV) typically varying from 120º to 150º, and the respective processing software are imperative for safe operation of autonomous vehicles (AVs). Characterizing the performance of sensing to determine the location of the host vehicle relative to other moving vehicles, all in path targets and road lane boundaries are equally important. Human driver sensing reports that approximately 90% of the primary factors behind crashes are due to human errors (see the statistics of the National Highway Traffic Safety Administration (NHTSA) in 2012). [6],[7] AVs with a high performance sensor suite are predicted to reduce crashes and injury rates by upwards of 50% as compared to non-AVs. Vehicle manufactures are motivated to contribute to transformative and highly beneficial technologies to support Intelligent Transportation Systems (ITS) and the future of the mobility industry. The state of the art reports the rationale for selection of IR sensors rather than ultrasonic sensors, RADARs, or cameras to perform the sensing function. This research proceeds with a conceptual design, implementation, testing, and verification of an IR sensor set interfaced to a single board computer. A desired outcome is to follow a vehicle that is remotely controlled. This vehicle is denoted as RC (remote controlled) down scaled-model vehicle. In the literature, Kou et al. employed two low cost sensors using a kinematic model of a car-like mobile robot (CLMR). [8] Chao-Lin Kuok et al., devised to compare two non-linear model-based approaches for autonomous vehicles. [9] Carson et al., and Englund et al. studied safety criteria for driver assisted systems and enabling technologies. [10],[11] Paden et al. investigated a decision-making hierarchy (i.e., route planning, behavioral decision making, motion planning, and vehicle control) for driverless vehicles. This paper is organized as follows. Section II describes the overall system configuration and sensor interfaces to accomplish the autonomous tracking capability. Here, we also present the embedded system design. Section III is devoted to the testing and the IR sensor configuration. We then describe analysis of the measured test results in Section IV. Section V provides a discussion and the conclusion.

2. Formulation of Research Project

This section describes the design and implementation of an embedded system onto the scale of 1/10 of remote controlled (RC) autonomous vehicles. The system design for this embedded system began by conducting tradeoff study between the Arduino and Raspberry Pi single board computers. In view of ease of programming, a large suite software algorithm readily available in the open source domain. Hence, the decision we have chosen the Raspberry Pi Model B+. It is a fully featured very compact computer operating at 700 MHz and uses the Raspbian Operating System (OS). The computer interface circuitry has mainly 7 components; namely,

- 4 Universal Serial Bus (USB) 2.0 slots
- Single storage data card slot
- 24 Pin Header
- Ethernet port
- Customer Support Identifier (CSI) connector for a camera
- High-Definition Multimedia Interface (HDMI) output
- Power from micro-USB Connector

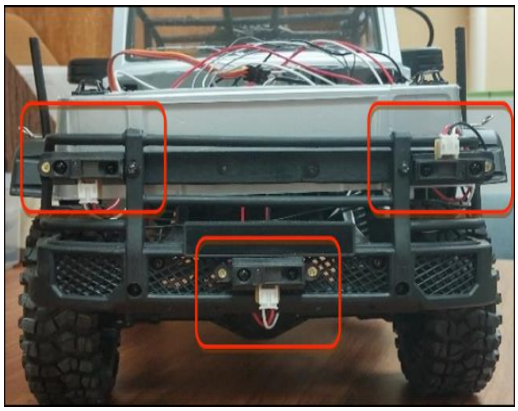


Figure 2. Mounting of IR sensors

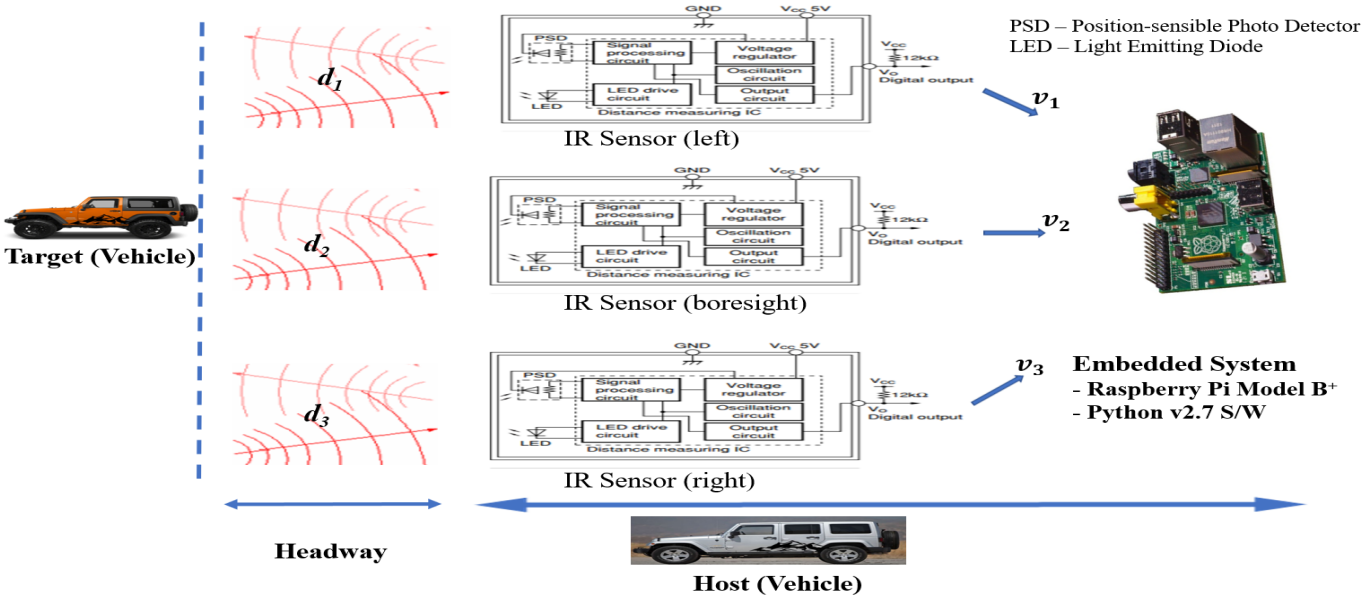


Figure 3. Autonomous convoy embedded system design. [16]

Since cost is always a consideration, an inexpensive SharpGP2YOD21YK IR Sensor operating at the wavelength λ=780 nano-meter (nm) is selected. [12]. The next step is to determine the optimal Field of View (FOV) for forward coverage on the model robotic vehicle. Figure 2. Depicts the mounting of three IR sensors. Specifically, the sensors are mounted in the headlight areas and lower front center of the RC controlled model. Work proceeds to install, calibrate and independently test the sensor in a standalone modes to consume less energy and to operate with nonlinear characteristics [9].

The relationship between the distance d_i and the operating voltage is inversely proportional, and it is given as

$$d_i = k\frac{1}{v_i^a}, \quad i = 1,2,3. \tag{1}$$

with k being the scaling constant and a is the exponent of proportionality. Equation (1) is empirical formula and for the given Sharp GP2YOA21YK IR sensor, it is specified as

$$d = \left(\frac{1}{c * A_{DC} + b}\right) - K \tag{2}$$

where distance d in cm, K is a corrective constant, A_{DC} is the digitized output voltage, b and c are variable constants to be evaluated from trend line equation. Besides proposed equation (2), the governing sensor equation (3) has three variables (b, c, and K) which are evaluated from the measured data.

The embedded Time-Division Multiplexing (TDM) system is implemented by discrete electronic components mounted on a Printed Circuit Board (PCB). The hardware is physically mounted into the body of the robotic model vehicle. Also, it is evident that the data streaming from sensors are converted into a single signal by segmenting the signal with short durations. Figure 3 shows the electrical interface and message traffic block diagram of the Sharp IR sensor in which the digital output V_o is directed to the digital I/O channel of a single board computer (i.e., Raspberry Pi Model B^+) [14][15]. This prototype design minimizes the sensor wiring and power connections.

The Python 2.7 language is employed to write the IR sensing system software illustrated in Figure 3. The algorithm is designed using classical Proportional Integral Derivative (PID) controller techniques. Specifically, the PI algorithm is selected as automatic, greater, and basic [17] to implement. SP is defined as a headway or fixed distance and independent of the measurement sensor PVs are the computed distances, shown in Figure 3. "PIDSYS" a MATLAB R2017A based function is utilized to return the parallel form of a continuous-time PI controller shown in Figure 4.

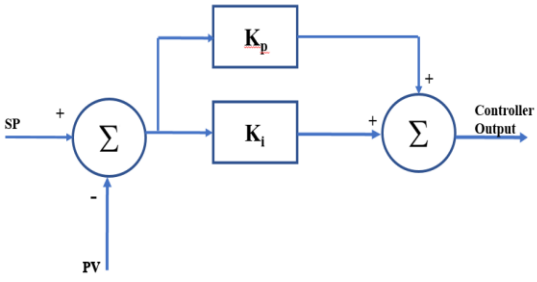


Figure 4. Block Diagram for PI controller.

Here, the instances of error are corrected by proportional term ($\mathbf{K_p}$) corrects and the accumulation of error is corrected by integral term ($\mathbf{K_i}$) corrects the. Based on the rise and fall of error signal {e(t)}, the amount added to the Controller Output (CtrlOut) increases or decreases immediately and proportionately. The CtrlOut is governed by Equation 3 and is given by

$$\text{CtrlOut} = \boldsymbol{K_p}\, e + \boldsymbol{K_i}\int e(\tau)\, d\tau \tag{3}$$

where the range of integration is [0, t], and e is error e = SP-PV required to achieve the controller output. The host vehicle track was based on the sensor voltages and the distances computed, independent of sensor specifications, that occurs per Equation 4.

$$d_1 > (d_2 + d_3) \Rightarrow \text{host tracks target to left} \tag{4a}$$

$$d_1 \simeq d_2 \simeq d_3 \Rightarrow \text{host tracks target straight} \tag{4b}$$

$$d_3 > (d_1 + d_2) \Rightarrow \text{host tracks target to right} \tag{4c}$$

The last stage of this research project formulation is to test the PI algorithm and ensure it is properly tuned or adjusted by starting with low proportional and no integral. The values of K_p and K_i are returned as 1.14 and 0.454, respectively. Figure 5 presents the deviation error of measured PV from SP in response to the computation of step response for this PI controller.

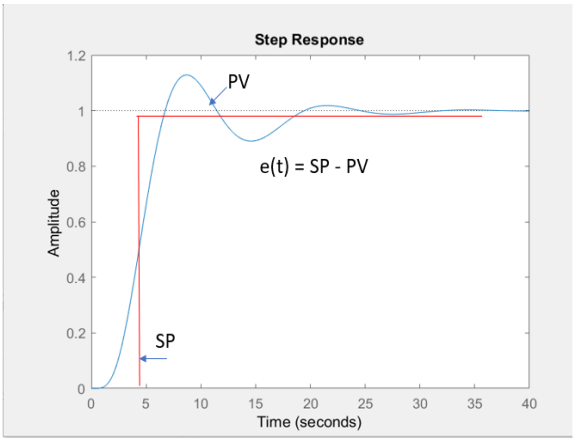


Figure 5. Step response for AV PI controller

3. Analysis of Test Results

After some troubleshooting and software debugging, smooth hallway surfaces are used to perform the initial simulation trials with the embedded system on the host vehicle following a lead vehicle where a smooth following is demonstrated for several test conditions. Figure 6 depicts static photos of the host vehicle chasing the target vehicle in the actual demonstration of this embedded system. The IR sensors are use in individual configurations (i.e., degraded FOV) or by combining that which addresses the redundancy of a given number of sensors



Figure 6. Host tracks target as per equation (4)

4. Data Fusion and Simulation

Sensor fusion is the most important and critical process for autonomous vehicles on the road for increased safety. The short comings of individual sensor can be overcome by fusing the data in such a way that each sensor can complement or augment in the presence of the other sensor failure. The failure or the malfunctioning of the sensors installed could cause by a natural phenomenon or manmade phenomenon [18]. In this parlance, the adaption of the sensor fusion methodology can maintain the essential functions for a required level of safety. With these insights, this section presents the advantage of using Kalman filter-based data fusion over single IR sensor and amid individual sensor failure. Equation 2 is utilized for range tracking and a complete mathematical modeling and simulation based on optimal state estimation theory is presented. With assumed and known values of measurement errors caused by system noises, an optimal estimator computes and processes the measurements for reducing the state error estimate using measurement system dynamics [19].

A. System Dynamic Model

The system dynamic model is a generic representation given by

$$\hat{X}_k(-) = F\hat{X}_{k-1}(+) + Gw_{k-1} \quad (5)\ w_k \sim N(0, Q_k)$$

$$Z_k = H\hat{X}_{k-1}(+) + v_k \quad (6)\ V_k \sim N(0, R_k)$$

where $\hat{X}_{k-1}(+)$ is a priority or the initial estimate of $\hat{X}_k$ taken as [6,1,1] for the purpose of this simulation. F or ϕ is denoted as the state transition matrix for predicting the future states and is given as

$$\phi = \begin{bmatrix} 1 & T & T^2/2 \\ 0 & 1 & T \\ 0 & 0 & 1 \end{bmatrix} \quad (7)$$

Here, the process noise w_k and the measurement noise v_k are modeled as white Gaussian noise with mean of zero and standard deviation of one. The noises in IR sensor and systems are described as statistical fluctuations or distortions in electrical current modeled by various mathematical models as [20]

$$\text{Photovoltaic noises or } \sigma_{\text{noise}} = \sqrt{\text{jn(r)} + \text{sn(r)} + \text{pn(r)}} \quad (8)$$

$$\text{Photoconductive noises or } \sigma_{\text{noise}} = \sqrt{\text{jn(r)} + \text{grn(r)} + \text{pn(r)}} \quad (9)$$

$$\text{where} \quad \text{jn(r)is Johnson noise} = \frac{4kT_dB(R)}{R_d} \quad (10)$$

$$\text{sn(r)is Dark current shot noise} = 2qi_DB(R) \quad (11)$$

$$\text{grn(r)is Generation Recombine noise} = 4qGi_DB(R) \quad (12)$$

$$\text{pn(r) is Photon noise} = \frac{2\eta q^2(P_s+P_b)B\lambda}{hc} \quad (13)$$

For a particular sensor and with its specifications the measurement noise variances can be modeled by using the equations from 12 to 17.

B. Fusion Algorithm

For the purpose of simulation, the sensor and the system are considered as time invariant systems in which all three IR sensors produce measurement at the same time. The method of the fusion process [21] depend on the covariance matrix and its trace operation, given by

$$\hat{x}_k = \sum_{i=1}^{N} a_i z_i(k) \quad (14)$$

where N is the total number of IR sensors mounted on the convoy, z(k) denotes the sensor value from the ith IR sensor for every time interval k, and the quotient a_i is given by

$$a_i = \frac{1/tr((R_i(k))}{\sum_{J=1}^{N}\frac{1}{tr((R_j(K))}} \quad (15)$$

where *tr()* performs the trace operation with mapped measurement covariance $R_i(k)$ of the ith IR sensor at k^{th} *time interval.* The quotient a_i is summed to unity by assumption.

C. Kalman Filter steps

Measurements from each sensor are fused using the below algorithm [21]:

- Perform the state estimate extrapolation expressed as
  $\hat{X}_k(-) = \phi_{k-1}\hat{X}_{k-1}(+)$ (16)
- Perform the error covariance extrapolation given as
  $P_k(-) = \phi_{k-1}P_{k-1}(+){\phi^T}_{k-1} + Q_{k-1}$ (17)
- Update the state estimate observation
  $\hat{X}_k(+) = \hat{X}_k(-) + \bar{K}_k[Z_k - H_k\hat{X}_k(-)]$ (18)
- Calculate the error covariance update
  $P_k(+) = [1 - \bar{K}_k H_k]P_k(-)$ (19)
- Calculate the Kalman Gain Matrix
  $\bar{K}_k = P_k(-){H_k}^T[H_kP_k(-){H_k}^T + R_k]^{-1}$ (20)

The bock diagram describing the system, measurement model, and discrete-time Kalman Filter is shown in Figure 7.

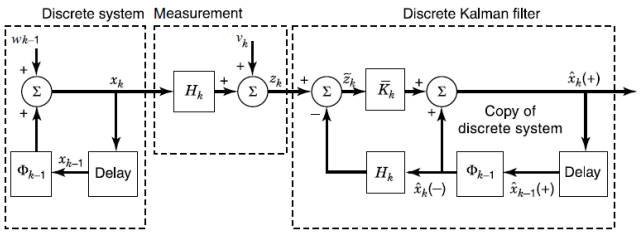


Figure 7. System model and Discrete Kalman filter [22]

The initial values for predicted covariance and state transition matrix are taken as

$$P_{k-1}=\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \quad (21)$$

Figure 7 presents the iterative recursion process of the Kalman filter and sums up the equations described in 17 to 21. The simulation results shown in the Figures 9 and 10 describe the effectiveness of Kalman filter-based data fusion. These iterative recursive filters are used for continuous time problems and considered as a breakthrough for estimation in linear dynamic systems. The propagation of covariance matrix and the dynamic calculation of the Kalman filter gains make this filter superior. The state vector dimension is 3-by-1 known as one dimensional third order filter tracking with one position or relative distance component, one velocity component and one acceleration component obtained by each individual IR sensor. The measurements from each sensor are fused using the algorithm and the fused measurement is optimally estimated by the filter steps mentioned in equations 17 to 21. Next, using the state estimate and the covariance matrix the propagation of prediction and correction will continue to obtain the smoothed estimate.

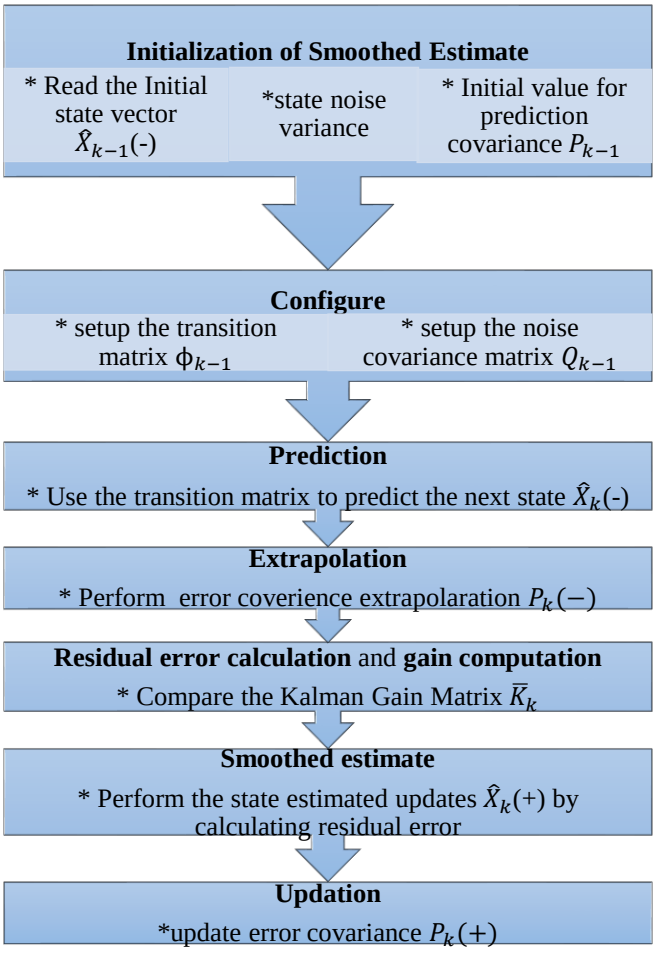


Figure 8. Iteration steps for Kalman filter [23]

Figures 8 presents the MATLAB simulation flowchart. Figure 11 indicates the digitally computed true distance between the target vehicle and the sensor mounted vehicle simulated based on the equation 2. It also shows the measured distances d1, d2 and d3 from each IR sensor and fused distance estimate based on Kalman filter.

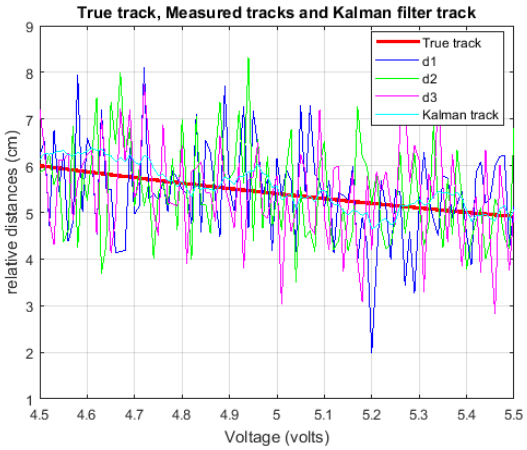


Figure 9. Tracks simulation

The simulation parameters are chosen such as sampling interval of 0.1 sec over the supply voltage values between 4.5 volts to 5.5 volts. The initial values of relative distances d1, d2 and d3 are 6.2, 5.8 and 7.1 cm. These values of relative distances are chosen based on the distance measuring range of the sensor considered in section 2. For each sample the values of d1, d2 and d3 vary between minimum of 4 cm to a maximum of 7 cm. Also, the minimum and the maximum values of true values of track is 4.9 cm to 6cm, and from Figure 10 it is evident that Kalman filter track has the best estimate that follows the true track closely. Table 1 shows the RMS errors and the performance of each sensor and the Kalman filter is analyzed by computing the distance error in relative position between the two vehicles using

$$RMS\ Error_{dist} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(x_{true}-x_{esti})^2}{3}} \quad (22)$$

In equation 22, x_{rue} are the true track values and x_{esti} are the individual estimates of each sensor and Kalman filter estimates as well. Therefore, using the equation 22, the RMS error values for each sensor tracks alongside the Kalman filter based fusion tracks are obtained and presented in Table 1.

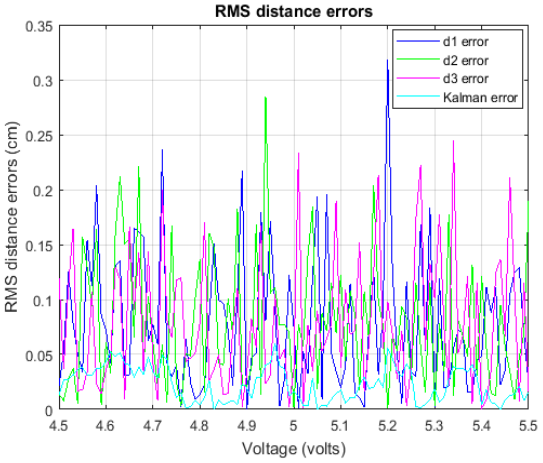


Figure 10. RMS errors in relative distances

From Figure 10 and Table 1, it is evident that Kalman filter based fusion method shows the better performance compared to individual IR sensors and the superiority and the advantage of implementing optimal estimation based fusion process.

Table 1. Average RMS Errors

Estimator	RMS Error (cm)
IR sensor 1	0.0230
IR sensor 2	0.1893
IR sensor 3	0.016
Kalman filter based fusion	**0.015**

Conclusion

This research project resulted in the successful development and implementation of an embedded system design for an autonomous vehicle. The implementation based on theoretical multisensory data fusion is based upon IR sensors and a single board computer. A dynamic demonstration of active following by the host is successful and was accomplished while considering the design constraints of the embedded system. The performance achieved on the 1:10 scaled model indicates positive proof of concept for real-world scaling of OEM (Original Equipment Manufacturer) vehicles. The system architecture provided for sensors from varied range of specifications to be incorporated and demonstrate performance of maintaining a set headway and smoothly following a target vehicle. From the test results, an overall error is found to be less than 5% in following distance and a measurement order of eight seconds resulted in response to steady state. A comparison to other research results reported in the literature confirms similar result where the trade of is use of two sensors (i.e., narrower FOV) whereas this work utilized a configuration of three sensors ((i.e., broader FOV). Also, the incorporation of the fusion applications for the robustness of distance and tracking, could potentially improve the safety considerations. Research results included determination of the system RMS error based on Kalman filter which provides greater accuracy than other computational approaches for error calculations. As error is reduced, AV operational safety will be enhanced. The average RMS error for the lower IR sensor (i.e., centered position and single sensor configuration) is within +/- .001cm. Furthermore, the contextual practical application of the sensor fusion for automotive applications reliability detects in-path targets that aid in collision avoidance. This research maybe be extended by configuring dissimilar sensors such as RADAR, LiDAR in conjunction with IR sensors leading to an Advanced Driver Assistance System (ADAS). These systems are totally dependent on weighting sensor information to make intelligent and safe decisions. Based on the theoretical simulation of sensor fusion presented in the paper, it is believed the paradigm of data fusion can enhance the reliability and robustness of an ADAS for improved safety considerations. The future scope of this research will be demonstrated in the real-world environment amid poor visible conditions to check the accuracy of the proposed methods outlined above. An approach may include resolving distances based on the speed of light by measuring the Time-of-Flight (ToF) between the sensor and the targeted image using a ToF Camera or LIDAR. Motion planning conducted by a low-level feedback controller is also a consideration for additional research since trajectory and path planning are computationally complex and require computing resources well beyond the selected Raspberry Pi Model B^+ single board computer.

References

[1] N. Lang *et al*., "*Making Autonomous Vehicles a Reality: Lessons from Boston and Beyond*., The Boston Consulting Group, Oct. 2017.

[2] A. Jonas *et al.*, "Autonomous Cars: Self-Driving the New Auto Industry Paradigm," *Morgan Stanley Blue Pap.*, pp. 1–109, 2013.

[3] L. Gannes, "Google Introduces New Self Driving Car at the Code Conference," *Re/code*. 2014.

[4] L. Mearian, "Nissan plans to offer affordable self-driving cars by 2020," *Computerworld*. 2013.

[5] D.A. Dickmanns and D. L. Wiesenthal, "Traffic congestion, driver stress, ana driver aggression," Agbressive behavior, pp 409 - 423, 1988.

[6] NHTSA, "U.S. Department of Transportation releases policy on automated vehicle development," *Nhtsa 14-13*, 2013.

[7] C. Pozna and C. Antonya, "Issues about autonomous cars," in *SACI 2016 - 11th IEEE International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, 2016.

[8] C. L. Kuo, N. S. Pai, Y. P. Kuo, and Y. C. Hu, "Following method for a car-like mobile robot using two IR sensors," in *IEEE International Conference on Control and Automation, ICCA*, 2014.

[9] E. Alcal; *et al.*, "Comparison of two non-linear model-based control strategies for autonomous vehicles," *2016 24th Mediterranean Conference on Control and Automation (MED)*. 2016.

[10] O. M. G. Carsten and L. Nilsson, "Safety assessment of driver assistance systems," *Eur. J. Transp. Infrastruct. Res.*, 2001.

[11] C. Englund *et al.*, "Enabling technologies for road vehicle automation," in *Road Vehicle Automation 4*, 2017.

[12] Sharp Corporation, "Distance Measuring Sensor GP2YOA21YK," 2016.

[13] B. Paden et .al, "A Survey of Motion Planning and Control Techniques for Self-driving Urban Vehicles., IEEE Transactions on Intelligent Vehicles, pp. 33-55. Volume: 1, Issue: 1, March 2016.

[14] Raspberry Pi Foundation, "Raspberry Pi - Teach, Learn, and Make with Raspberry Pi," w*ww.raspberrypi.org*, 2012.

[15] E. Upton and H. Gareth, *Raspberry Pi User Guide*. 2014.

[16] D. Gunnarsson, S. Kuntz, G. Farrall, A. Iwai, and R. Ernst, "Trends in Automotive Embedded Systems," in *CASES'12: Proceedings of the 2012 ACM International Conference on Compiliers, Architectures and Synthesis for Embedded Systems, 2012*.

[17] R.-E. Precup, A.-D. Balint, E. M. Petriu, M.-B. Radac, and E.-I. Voisan, "PI and PID controller tuning for an automotive application using backtracking search optimization algorithms," *SACI 2015 - 10th Jubil. IEEE Int. Symp. Appl. Comput. Intell. Informatics, Proc.*, 2015.

[18] Hannes Estl, "Sensor Fusion: A Critical step on the road to autonomous vehicles", Automotive ADAS sector, April 11th 2016.

[19] A. Gelp et al, "*Applied optimal estimation*", Second edition, MIT press, February 1974.

[20] Christopher W. Keefer, "Infrared Target Detection: Signal and Noise Sensitivity Analysis", Master's Thesis, December 1989, School of Engineering of the Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio.

[21] X. Yu, J. Yihui, Z. Yan, "Several methods of radar data fusion", IEEE 3rd International Symposium on Electromagnetic Compatibility, May 2002, China.

[22] Mohinder S. Grewal and Angus P. Andrews, *Kalman filtering theory and practice using MATLAB*, 2 Edition, Wiley & Sons, Inc.

[23] David Kondru, Mohan Krishna, "Kalman filter based Target Tracking for Track while scan Data Processing", 2015 2nd IEEE International Conference on Electronics and Communication Systems (ICECS), 26-27 Feb. 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 382-393 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Virtual Watershed System: A Web-Service-Based Software Package For Environmental Modeling

Rui Wu[1], Connor Scully-Allison[*,2], Moinul Hossain Rifat[2], Jose Painumkal[2], Sergiu Dascalu[2], Jr. Frederick C Harris[2]

[1]*Department of Computer Science, East Carolina University, 27858, USA*

[2]*Computer Science & Engineering Department, University of Nevada, Reno, 89557, USA*

ARTICLEINFO

Article history:
Received: 31 July, 2018
Accepted: 07 October, 2018
Online: 22 October, 2018

Keywords:
Web-based Virtual Watershed System
Model as service
Cloud-based application
Hydrologic model

ABSTRACT

The physically-based environmental model is a crucial tool used in many scientific inquiries. With physical modeling, different models are used to simulate real world phenomena and most environmental scientists use their own devices to execute the models. A complex simulation can be time-consuming with limited computing power. Also, sharing a scientific model with other researchers can difficult, which means the same model is rebuilt multiple times for similar problems. A web-service-based framework to expose models as services is proposed in this paper to address these problems. The main functions of the framework include model executions in cloud environments, NetCDF file format transmission, model resource management with various web services. As proof of concept, a prototype is introduced,implemented and compared against existing similar software packages. Through a feature comparison with equivalent software, we demonstrate that the Virtual Watershed System (VWS) provides superior customization through its APIs. We also indicate that the VWS uniquely provides friendly, usable UIs enabling researchers to execute models.

1 Introduction

Modeling has become an indispensable tool in growing environmental scientists' understanding of how natural systems react to changing conditions. It sheds light on complex environmental mysteries and helps researchers in formulating policies and decisions on future scenarios. Environmental modeling is highly challenging as it involves complex mathematical computations, rigorous data processing, and convoluted correlations between numerous parameters. Three commonly-used and important scientific software quality measures are maintainability, quality, and scalability. Issues like data storage, coupling models, retrieval, and running are hard problems and need to be addressed with extra efforts from software engineering perspective. It is a challenging job to design integrated systems that can address all these issues.

It is essential to build high-quality software tools and design efficient frameworks for scientific research. Abundant scientific model data are generated and collected in recent years. Software engineering can assist this emergence through the creation of distributed software systems and frameworks enabling scientific collaboration with previously disparate data and models. It is challenging to implement software tools for interdisciplinary research because of the problem of communication and team building among different scientific communities. For example, the same terminology can have different meanings in different domains. This increases difficulties in comprehending software requirements when a project involves stakeholders (e.g. researchers) from different fields.

Most work described in this paper is for Watershed Analysis, Visualization, and Exploration (WC-WAVE), which is a NSF EPSCoR-supported project and initiated by jurisdictions of EPSCoR of Nevada, Idaho and New Mexico. The WC-WAVE project includes three principal components: watershed science, data cyberinfrastructure and data visualization [1]. The project main goal is to implement VWS with the collaborations

[*]Connor Scully-Allison, 1664 N. Virginia Street, CSE (0171), Reno, NV 89557, (775) 771-1469 & cscully-allison@nevada.unr.edu

https://dx.doi.org/10.25046/aj030544

between cyberinfrastructure team members and hydrologists. The platform is able to store, share, model and visualize data on-demand through an integrated system. These features are crucial for hydrologic research.

Different hydrologic models, such as ISNOBAL and PRMS, are commonly used by WC-WAVE hydrologists. These models are leveraged to predict or examine hydrologic processes of Lehman Creek in Nevada, Dry Creek and Reynolds Creek in Idaho, and Jemez Creek in New Mexico. We propose a framework for representing model data in a standardized format called the "Network Common Data Format" (NetCDF) and exposing these hydrologic models through web services. This framework is based on our previous work introduced in [2]. To improve on our prior work we have implemented some new Docker APIs to control system components wrapped in docker containers. To simplify data extraction, modification, and storage, NetCDF data format [3] is used in the system for grid-organized scientific data. Most parts of the system framework can be reused for data-intensive purpose because it is designed with the blueprint and template concepts. ISNOBAL and PRMS are physically-based models which can produce very accurate results. However, these two models require abundant computing power. To solve the challenge, we leverage a cluster to execute models in parallel. ISNOBAL and PRMS are used in this paper to demonstrate the ideas and functionality of the proposed framework. Throughout the remainder of this paper, we refer to this prototype system as the VWS.

ISNOBAL is a grid-based DEM (Digital Elevation Model) and created to model the seasonal snow cover melting and development. The model author is Marks et al. [4] and it is initially developed for Utah, California, and Idaho mountain basins. The model determines runoff and snowmelt based on terrain, precipitation, region characteristics, climate, and snow properties [4].

PRMS is short for Precipitation-Runoff Modeling System and is initially written with FORTRAN in 1983. PRMS is prevalent physical process based distributed-parameter hydrologic model and the main function of a PRMS model is to evaluate a watershed response to different climate and land usage cases [5, 6, 7]. It composed of algorithms describing various physical processes as subroutines. The model, now in its fourth version, has become more mature over the years of development. Different hydrology applications, such as measurement of groundwater and surface water interaction, the interaction of climate and atmosphere with surface water, water and natural resource management, have been done with the PRMS model [5, 6, 7].

This paper is organized as follows in its remaining sections: Section 2 introduces background and related work; Section 3 describes the system design; Section 4 describes the prototype system and how the software was built using RESTful APIs; Section 5 compares our work with related tools; and Section 6 contains the papers conclusions and outlines planned future work.

2 Background and Related Work

"How to implement software for interdisciplinary research?" is an interesting question and there exists some successful work on environments and frameworks which seek to answet this question. In this section, relevant, popular earth science applications and frameworks are introduced.

Community Surface Dynamics Modeling System (CSDMS) was a project started in 1999 to conduct expeditious research of earth surface modelers by creating a community driven software platform. CSDMS applies a component-based software engineering approach in the integration of plug-and-play components, as the development of complex scientific modeling system requires the coupling of multiple, independently developed models [8]. CSDMS allows users to write their components in any popular language. Also they can use components created by others in the community for their simulations. CSDMS treats components as pre-compiled units which can be replaced, added to, or deleted from an application at runtime via dynamic linking. Many key requirements drove the design of CSDMS, including the support for multiple operating systems, language interoperability across both procedural and object-oriented programming languages, platform independent graphical user interfaces, use of established software standards, interoperability with other coupling frameworks and use of HPC tools to integrate parallel tools and models into the ecosystem.

A leading hydrologic research organization is CUAHSI. "CUAHSI" , and acronym for "Consortium of Universities for the Advancement of Hydrologic Science Inc.," represents universities and international water science-related organizations. One of most highly esteemed products is HydroShare, which is a hydrologic data and model sharing web application. Hydrologists can easily access different model datasets and share their own data. Besides this, this platform offers many distributed data analysis tools. A model instance can be deployed in a grid, cloud or high-performance computing cluster with HydroShare. Also, a hydrologist is able to publish outcomes of their research, such as a dataset or a model. In this way, scientists use the system as a collaboration platform for sharing information. HydroShare exposes its functionality with Application Programming Interfaces (APIs), which means its web application interface layer and service layer are separated. This enables interoperability with other systems and direct client access [9].

Model as a Service is proposed by Li et al. [10] for Geoscience Modeling. It is a cloud-based solution and [10] has implemented a prototype system to execute high CPU and memory usage models remotely as a service with third party platform, such as AWS (Amazon Web Service) and Microsoft Azure.

The key idea of MaaS is that model executions can be done through a web interface with user inputs. Computer resources are provisioned with a cloud provider, such as Microsoft Azure. The model registration is done in the framework with a virtual machine image

repository. If a model is registered and placed in the repository, it can be shared by other users and multiple model instances can be executed in parallel based on demand.

McGuire and Roberge designed a social network to promote collaboration between watershed scientists. Despite being highly available, the collaboration between the general public, scientists, and citizen has not been leveraged. Also, hydrologic data is not integrated in any system. The main goal of this work is to design a collaborative social network for multiple watershed scientific and hydrologic user groups [11]. However, more efforts need to be done.

The Demeter Framework by Fritzinger et al. [12] represents another attempt to utilize software frameworks as a scientific aid in the area of climate change research. A software framework named "The Demeter Framework" is introduced in the paper and one of the key ideas is a component-based approach to integrate different components into the system for the "model coupling problem." "The model coupling problem" refers to using a model's outputs as another model's inputs to solve a problem.

Walker and Chapra proposed a web-based client-server approach for solving the problem of environmental modeling compared to the traditional desktop-based approach. The authors assert that, with the improvement in modern day web browsers, client-side approaches offer improved user interfaces compared to traditional desktop software. In addition, powerful servers enable users to perform simulations and visualizations within the browser [13].

The University of New Mexico has implemented a data engine named GSToRE (Geographic Storage, Transformation and Retrieval Engine). The engine is designed for earth scientific research and the main functions of the engine are data delivery, documentation, and discovery. It follows the combination of community and open standards and implemented based on service oriented architecture. [14].

2.1 Service Oriented Architecture

Industry has shown more and more interests on Service Oriented Architecture (SOA) to implement software systems. [15]. The main idea of SOA is to have business logic decomposed into different units (or services). These units are self-contained and can be easily deployed with container techniques, such as Docker.

Representational State Transfer Protocol (REST) is primarily an architectural style for distributed hypermedia systems introduced by Fielding, Roy Thomas in his Ph.D. dissertation [16]. REST defines a way for a client-server architecture on how a client and a server should interact, by using a set of principles. REST has been adopted for building the main architecture of the proposed system. Statelessness, uniform Interface, and cache are the main characteristics of a REST client-server architecture [16]. It is for these characteristics that REST is leveraged in our proposed system.

Statelessness Statelessness is the most important property or constraint for a client-server architecture to be RESTful. The communication between the client and the server must be stateless, which means the server is not responsible for keeping the state of the communication. It is the client's responsibility. A request from the client must contain all the necessary information for the server to understand the request [16, 17]. Two subsequent requests to the server will not have any interdependence between each other. Introducing this property on the client-server architecture presents several benefits regarding visibility, reliability, and scalability [16, 17]. For example, as the server is not responsible for keeping the state and two subsequent requests are not interrelated, multiple servers can be distributed across a load-balanced system where different servers can be responsible for responding to different requests by a client.

Uniform Interface Another important property of a RESTful architecture is it provides a uniform interface for the client to interact with a server. Instead of an application's particular implementation, it forces the system to follow a standardized form. For example HTTP 1.1 which is a RESTful protocol provides a set of verbs (e.g., GET, POST, PUT, DELETE, etc.) for the client to communicate with the server. The verbs, such as "GET" and "POST", work as an interface making the client-server communication generic[16].

Cache REST architecture introduces cache constraint to improve network efficiency [16]. A server can allow a client to reuse data by enabling explicitly for labeling some data cacheable or non-cacheable. A server can serve data that will not change in the future as cached content, allowing the client to eliminate partial interaction with that data in a series of requests.

2.2 REST Components

The main components of REST architecture include resources, representations, and resource identifiers.

Resources The resource is the main abstract representation of data in REST architecture [16]. Any piece of data in a server can be represented as a resource to a client. A document, an image, data on today's weather, a social profile, *everything* is considered a resource in the server. Formally, a resource is a temporarily varying function of $M_R(t)$ that maps to a set of entities for time t [16].

Representations A resource is the abstract building block of the data in a web server. For a client to consume the resource it needs to be presented in a way the client can understand. This is called representation. A representation is a presentation format for representing the current state of a resource to a consumer. Some commonly used resource representation format in the current standard are HTML (Hypertext Markup

Language), XML (Extensible Markup Language), JSON (JavaScript Object Notation), etc. A server can expose data content in different representations so that consumer can access the resources through resource identifiers (discussed in Section 2.2) in the desired format.

Resource Identifiers A resource is uniquely identified through a resource identifier in a RESTful architecture. For example, in HTTP Uniform Resource Identifier (URI) is used to identify a resource in a server. A URI can be thought as the address of a resource in the server [18]. A resource identifier is the key for a client to access and manipulate a resource in the server.

2.3 Microservice Architecture

Software as a Service (SaaS), as a new software delivery architecture, has emerged to leverage the widely-used REST standard for processing in addition to data transfer. The main advantage of SaaS is that it does not require local installation and this is a significant IT trend based on industry analysis [19].

Similarly, "microservice" decomposes an application into small components (or services) and these components communicate with each other through APIs. "Microservice" is a solution to monolithic architecture relevant problems. [20]. Because of the "microservice" characteristics, an application can be easily scaled and the deployment risks have been reduced without interrupting other services.

Traditional monolithic applications are built as a single unit using a single language stack and often composed of three parts, a front end client, a back-end database and an application server sitting in the middle that contains the business logic. Here, the application server is a monolith that serves as a single executable. A monolithic application can be scaled horizontally by replicating the application server behind a load balancer to serve the clients at scale. The biggest issue with monolithic architecture is that, as the application grows, the deployment cycle becomes longer as a small change in the codebase requires the whole monolith to be rebuilt and deployed [20]. It leads to higher risk for maintenance as the application grows. These pitfalls have lead to the idea of decomposing the business capabilities of an application into self-contained services.

Being a relatively new idea, researchers have attempted to formalize a definition and characteristics of *Microservices*. [20] [20] has put together a few essential characteristics of a microservice architecture that are described in brief in the following sections.

Componentization via Services The most important characteristic of a microservice architecture is that service functionalities need to be componentized. Instead of thinking of components as libraries that use in-memory function calls for inter-component communication, we can think of components in terms out-of-process services that communicate over the network, quite often through a web service or a remote procedure call. A service has to be atomic, doing one thing and doing one thing well [21].

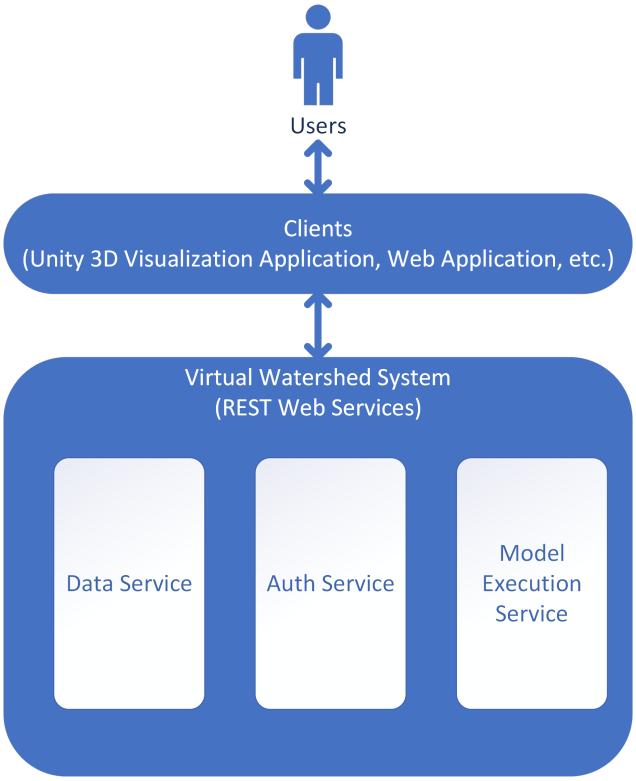


Figure 1: High-level diagram of VWS: clients and VWS communicate through REST Web Services. Possible clients include a script and applications. VWP provide data service, auth service, and model execution service.

Organized Around Business Capabilities Monolithic applications are typically organized around technology layers. For example, a typical multi-tiered application might be split logically into persistence layer, application layer and UI layer and teams are also organized around the technologies. This logical separation creates the need for inter-team communication even for a simple change. In microservices, the product is organized around business capabilities where a service is concentrated on one single business need and owned by a small team of cross-domain members.

Decentralized Governance In a monolithic application, the product is governed in a centralized manner, meaning it restricts the product to a specific platform or language stack. But in microservices, as each service is responsible implementing an independent business capability, it allows for building different services with different technologies. As a result, the team gets to choose the tools that are best suited for each of the services.

[21], in his book Building Microservices [21], has discussed several concrete benefits of using microservices over a monolith. Several important benefits are discussed below in brief:

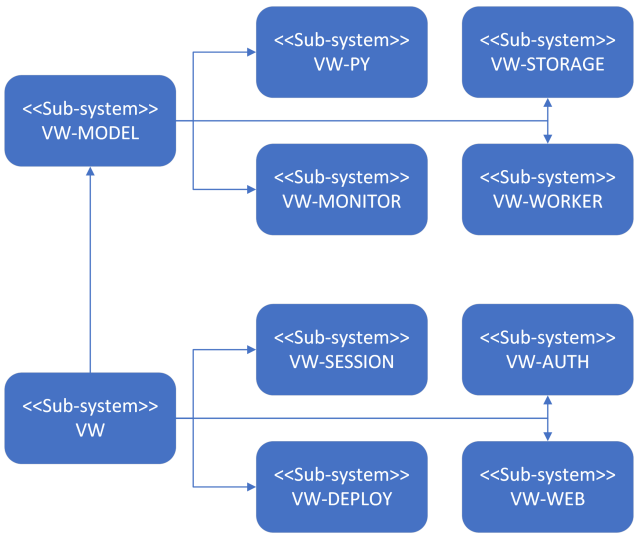


Figure 2: System-level diagram of VWS: two main system components are VW-MODEL and VW. VW-MODEL contains components relative to model execution and VW contains components relative to the platform.

Technology Heterogeneity When independent services are built separately, it allows adoption of different technology stacks for different services. This gives a team multiple advantages: liberty to choose a technology that best suits the need and adopt technology quickly for the needs.

Scaling Monolithic applications are hard to deal with when it comes to scaling. One big problem with monolithic applications is that everything needs to be scaled together as a piece. But microservice allows having control over the scaling application by allowing to scale the services independently.

Ease of Deployment Applying changes for a monolithic application requires the whole application to be re-deployed even for a minor change in the codebase. It poses a high risk as unsuccessful re-deployments even with minor changes can take down the software. Microservices, on the other hand, allow low-risk deployments without interrupting the rest of the services. They also allow having faster development process with small and incremental re-deployments.

3 Proposed Method

Here we introduce detailed design documentation for the VMS. We present the design using many common diagrams used in software engineering, including system diagrams and workflow diagrams. The VWS aims to create a software ecosystem named the *Virtual Watershed* by integrating cyberinfrastructure and visualization tools to advance watershed science research. Fig. 1 shows a general high-level diagram of components of the ecosystem. The envisioned system is centered around services comprised of data, modeling, and visualization components. A high-level description of each depicted component is provided in Fig. 1.

3.1 Data Service and Modeling Service

To create a scalable and maintainable ecosystem of services, a robust data backend is crucial. The envisioned data service exposes a RESTful web service to allow easy storage and management of watershed modeling data. It allows retrieval of data in various OGC (Open Geospatial Consortium) standards like WCS, WMS to allow OGC compliant clients to retrieve data automatically. This feature is very important for data-intensive hydrologic research and is essential for a scientific modeling tool.

Hydrologists often use different modeling tools to simulate and investigate the change of different hydrologic variables around watersheds. The modeling tools are often complex to setup in local environments and takes up a good amount of time setting up [10]. Besides these modeling tools may require high computational and storage resources that make them hard to run in local environments. The proposed modeling service aims to solve these issues by allowing users to submit model execution tasks through simple RESTful web service API. This approach of allowing model-runs through a generic API solves multiple problems:

- It allows the users to run models on demand without having to worry about setting up environments.
- It allows modelers to accelerate the process of running models with different input parameters by submitting multiple model-runs to be run in parallel which might not be feasible in local environment due to lack of computational and storage resources.
- It opens the door for other services and clients to take advantage of the API to automate the process of running models in their workflow.

The primary goal of this Virtual Watershed system is to allow watershed modelers to share their data resources and execute relevant models through web services without having to install the models locally. The system is designed as a collection of RESTful microservices that communicate internally. The web service sits on top of an extensible backend that allows easy integration of models and scalability over model runs and number of users connected to the service.

The system provides a mechanism for integrating new models by conforming with a simple event driven architecture that allows registering a model in the system by wrapping it with a schema driven adaptor.

As model execution is a CPU intensive process, scaling the server with growing number of parallel model execution is an important issue. The proposed architecture aims to solve this problem by introducing a simple database-oriented job queue that provides options for adding more machines as the system grows.

3.2 System Level Design

Several different submodules comprise The Virtual Watershed system. Each module provides different functionalities to the entire framework. This relationship between modules is described with Figure 2. The following paragraphs will explain each submodule in detail.

VW-PY: Using the VW-PY module, users can define adapters to configure compatibility between different models and the VWS. An adaptor is python code which encapsulates a model and that allows for it to be run programmatically. VW-PY handles many aspects of model "wrapping" with this python code through an interface. Through this API, an user can define the code which handles converting between data formats, executing models, and model-progress-based event triggering. This event triggering system provides model-wrapper developers with the tools to emit model execution progress.

VW-MODEL: The VW-MODEL submodule, through a web service, exposes a RESTful API to the user/client. Users can upload, query, and retrieve model run packages using this API.

VW-WORKER: The VW-WORKER is a service which encapsulates model adaptors in a worker service that is organized in a queue data structure. This component communicates with the VW-MODEL component using a redis data-backend.

VW-STORAGE: The VW-STORAGE module provides an interface for object storage for the VW-MODEL component. Developed as a generic wrapper, sysadmins can configure this interface to work with local or cloud-hosted storage providers.

VW-AUTH: The VW-AUTH module provides authentication services for users/clients connecting to the VWS framework. This service provides clients with a JWT token enabling them to securely utilize the other services described in this section.

VW-SESSION: The VW-SESSION sub module coordinates different components in the VWS to provide a common session backend for a single user. Data about each session is maintained and managed with a Redis data store shared between services.

VW-WEB: This is the web-application front-end which provides users with APIs to interact with the model processing modules described above. In a standard use case of this module, users will get access to a session after logging into this system with the VW-AUTH component. Once given a session, users can use this interface to access resources, run models, track progress and upload/download model run resources.

3.3 Detailed Design

The VWS is comprised of many distinct services and web applications which communicate with one-another to facilitate model runs. We are able to provide secure and centralized communications with the aid of a common authentication gateway. VW-AUTH was developed as a micro service to provide this functionality. It aggregates functionality for registration, authentication, and authorization for system users.

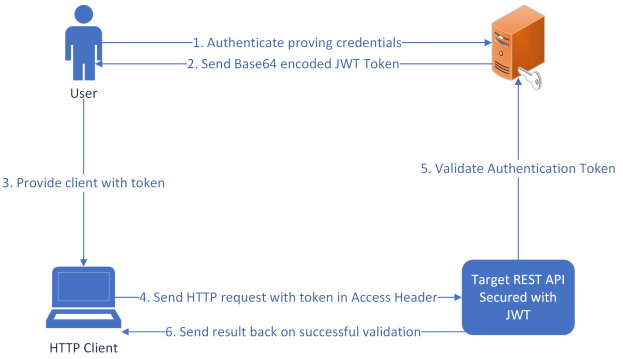


Figure 3: Workflow for accessing secure REST endpoint with JWT token

The Authentication component itself exposes RESTful endpoints that provide user access to different services, in addition to endpoints for authentication and registration. To enable the verification of a client's authentication by other services, a JSON Web Token (JWT) based authorization scheme is utilized. As an RFC standard for exchanging information securely between a client and a server, the using JWT ensures a high level of security in this system. A standard workflow for user authorization is depicted in Figure 3.

Users can manage uploaded models via a RESTful API endpoint provided by the Model Web Service. Using this endpoint, a properly authenticated user is able to request a model run, upload necessary input files needed by the model and start model execution. Model run data is stored by the VWS in a dedicated database located on a server operating out of the University of New Mexico [1].

Available models in the system are self-describing with schema that indicates necessary input files, formats, and execution policies. The schema also shows the mapping between user facing and model adaptor parameters. A user/client can use this schema to understand the necessary resources required to run a given model. The a typical model run from the perspective of the clinent side application involves six steps: 1) retrieve the server-side model schema; 2) instantiate a model run session on the server; 3) upload model inputs; 4) run the model; 5) track model run progress; and finally, 6) download outputs.

Figure 4 shows a simple workflow chart from a user or client's perspective. This workflow offers many advantages this workflow and sever-client architecture compared to traditional approaches to model running:

- Users do not need to know internal specifics of the model. Setup details and dependencies are hidden from the user, which provides a more seamless experience.
- Users don't need to worry about installing model dependencies.

- Users can easily initiate multiple parallel model runs on a server, expediting research.
- Users have ubiquitous access to server-side data via provided RESTful APIs.

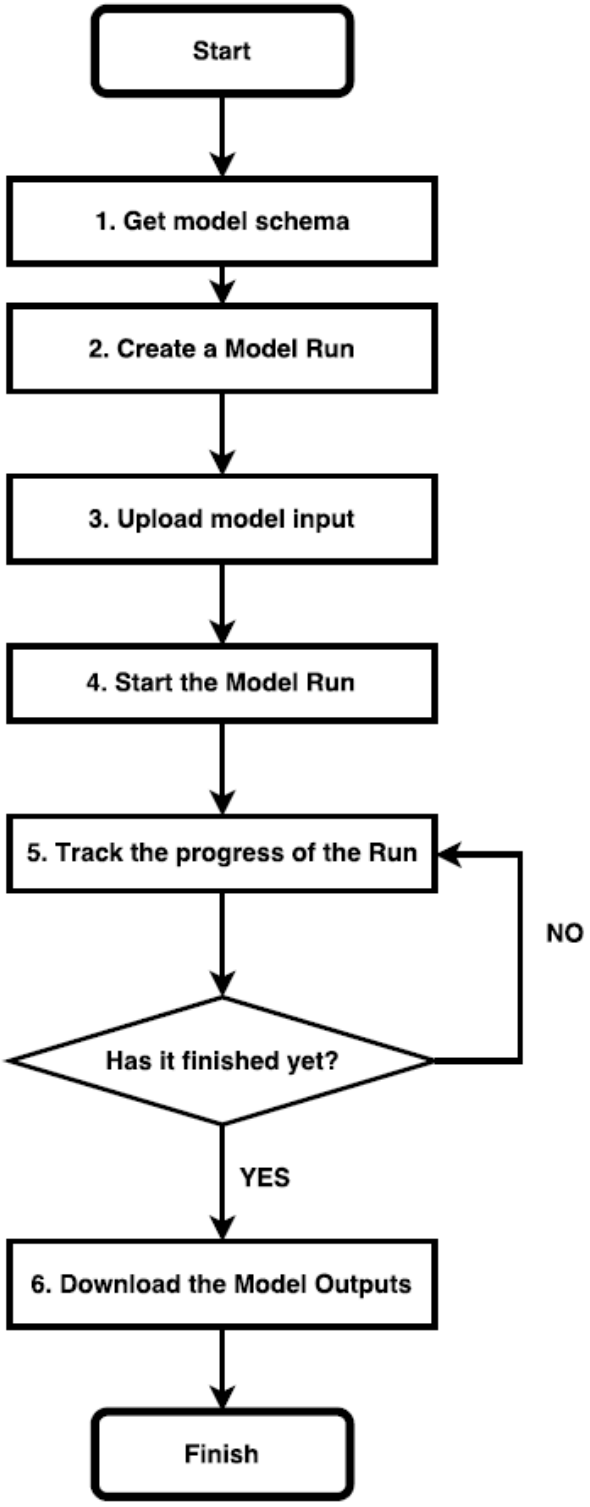


Figure 4: Workflow from user's perspective to run a model

The first step in creating the architecture for exposing "models as services" is enabling the programmatic running of models. A hurdle to accomplishing this is that a model can have many different dependencies required to run. Programmatic running is accomplished in this program by providing developers the tools to create python wrappers around models. These wrappers expose dependencies to model through container images.

In creating this system we had to strongly consider problems of data format heterogeneity with model inputs and outputs. Different models commonly accept and output data in different file formats. To achieve facilitate greater data interoperability between these models, we provided an option for developers to write NetCDF adapters for each models. A model adaptor is a Python program that handles data format conversion, model execution and progress notification. Adaptor developers must provide converters which define the method of conversion and deconversion between the native formats used by the model and netCDF. An execution function must also be implemented for the model. In this method, resource conversion and model execution occurs. The wrapper reports on model execution events via a provided "event emitter." An event listener can catch these events as they are reported by the emitter. This listener provides a bridge between the internal progress of the model and the user with the aid of a REST endpoint.

Through an adaptor, models can be encapsulated for programmatic execution. However, to bridge frontends with actual model execution a process is required from the model worker module. Utilizing a messaging queue, we create a bridge between the client frontend and worker backend. When a run task is submitted through from the client side, it is placed into the queue and assigned a unique id. The consumer/worker process listens to the queue through a common protocol for new jobs.

The worker service is a server-side python module that has access to a model's executable code and installed dependencies. The worker resides in an isolated server instance that contains the dependencies and libraries of the model installed. This module uses Linux containerization to facilitate easy deployment of model workers.

3.4 Deployment Workflow

A Linux-container-based deployment workflow has been devised for the for many different components of VWS. We utilized the containerization software, Docker, to enable our implementation of this workflow [22]. Docker containers have advantages over traditional virtual machines because they use fewer resources. This workflow allows for iterative deployment, simple scaling of the containerized components, and provides a strategy to register new models in the system.

Every VWS component is containerized with Docker. We have a set up a central repository of docker images which contain images for each component in this system. A docker "image" describes a template that provides Docker with the OS, dependencies of an

application and the application itself. Docker uses this template to build a working container. Each repository in the Virtual Watershed has a Dockerfile that describes how the component should be built and deployed into a query-able container. The repositories use webhooks to automate the building of docker containers.

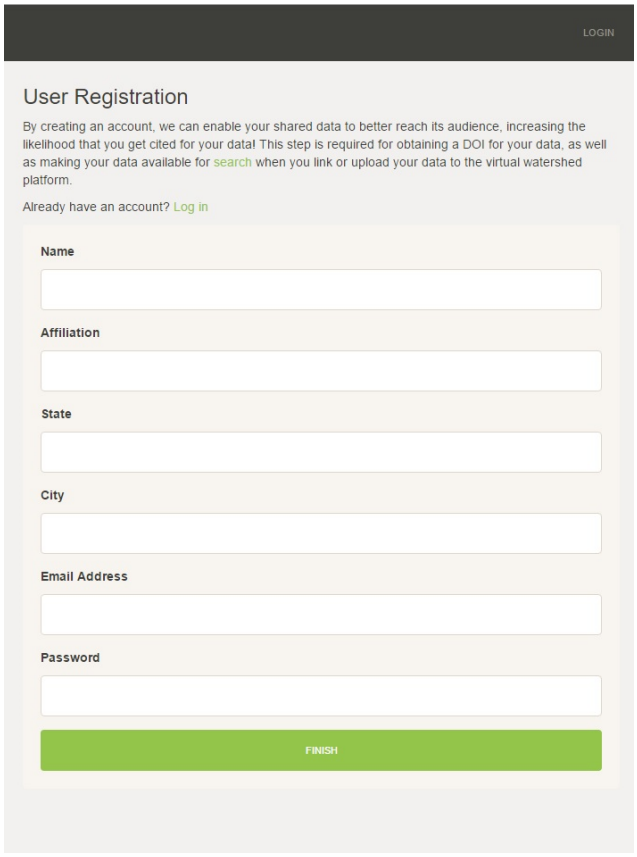


Figure 5: Registration page of Virtual Watershed System

It is easy to register new models in this system using this containerization workflow. Using our system, a user may register a model with the creation of docker image describing the model. With this image users can declare the OS, libraries and other dependencies for a model. A typical registration of a model requires the following steps: 1) create a repository for the model; 2) develop the wrapper; 3) specify dependencies within the dockerfile; and finally, 5) create a docker image in the image repository.

4 System Prototype and Testing

The project re-used some code and similar structures to those introduced in [23, 24]. The web service frontends were built using a Python micro-service framework called Flask, with various extensions [25]. Some key libraries used were: 1) Flask-Restless, used to implement the RESTful API endpoints; 2) SQLAlchemy, to map the python data objects with the database schema [26]; 3) PostgreSQL as the database [27]; and, 4) Flask-Security with Flask-JWT which provides the utilities used for security and authentication. Celery was used to implement the task queue for model workers [28]. Redis worked as a repository for model results output by Celery. A REST specification library called Swagger was used to create the specification for the REST APIs. For the web front end, HTML5, CSS, Bootstrap, and Javascript (with ReactsJS) were used.

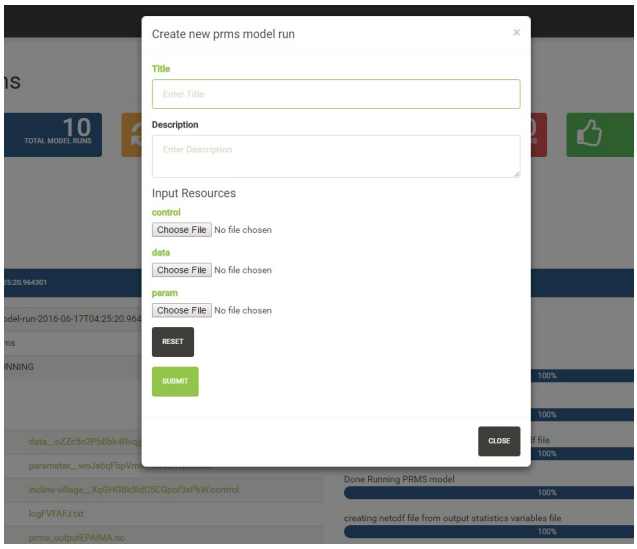


Figure 6: Dynamically generated upload form

To efficiently develop the VWS, an MVC design pattern was used to structure the code. The primary repository used to manage and distribute code was Github. Source code for this project is made publicly available through the Virtual Watersheds GitHub repository [28]. Dockerhub [29] is used for image management, which can automatically update images when Github code is changed.

The prototype system is comprised of two main components 1) the authentication module and 2) the modeling module. First, activities related to the VWS are handled by the authentication module. All standard user management functionality is handled by this module: registration, logins, verification, password management, and authentication token generation. Figure 5 shows the registration page of Virtual Watershed system. In addition to this interface, the VWS also provides endpoints for registration and authentication from user constructed scripts.

Modeling is necessary and commonly used in hydrologic research. Modification of the existing model simulations is a complex activity that hydrologists often have to deal with while analyzing complicated environmental scenarios. Modelers must make frequent modifications to underlying input files followed by lengthy re-runs. Programming languages could help significantly with this easily automated and repetitious task. However, it is complex and time consuming for hydrologists to write their own programs to handle file modifications for scenario based studies. To solve this challenge, the modeling module provides an in-

tuitive user interface where users can create, upload, run and delete models. The UI informs users about the progress of the model run with a bar that displays a percentage of completion. The module also includes a dashboard where users can view the models being run, the finished model runs, and also download the model run files. A user can execute multiple PRMS models in parallel by uploading the three input resources needed for PRMS model. The upload interface as shown in Figure 6 is generated dynamically from the model schema defined for each of the models.

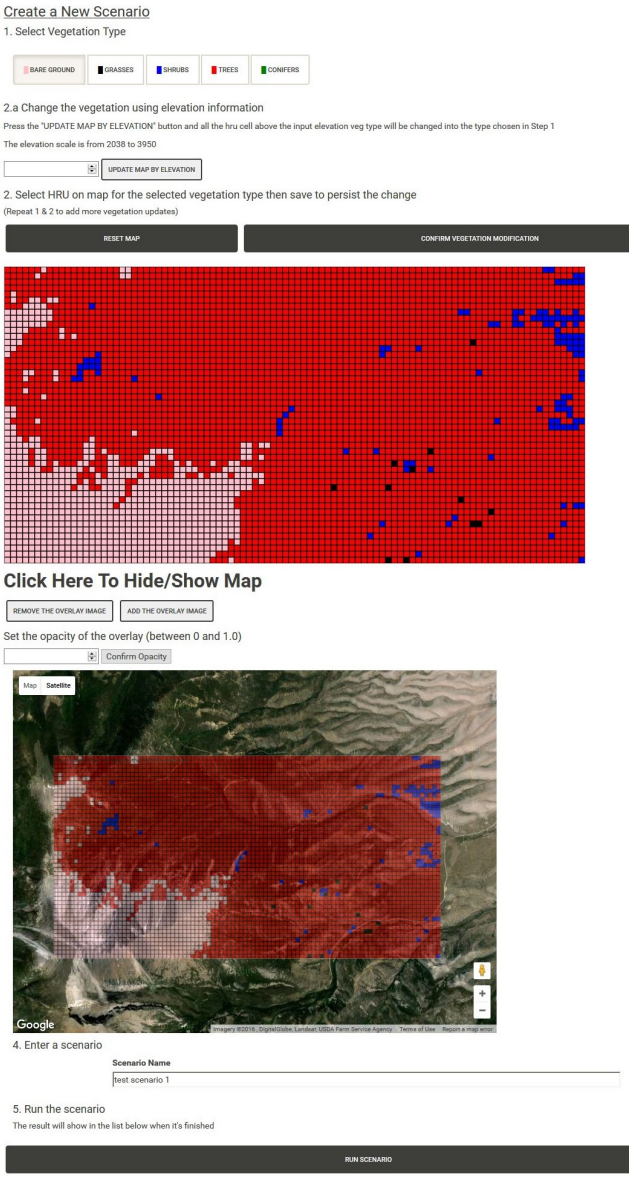


Figure 7: Scenario creation interface for PRMS model that uses Modeling REST API to execute model

The modeling interface displays peak utility when users use it to programmatically run a bundle of models in parallel without having to worry about resource management. In the prototype system, the client is used to adjust input parameters for models and for model execution.

Model calibration can be a time consuming process for many hydrologists. It often requires that the modeler re-run and re-run a model many time with slightly varied input parameters. For many modelers this is a manual process. They will edit input files with a text editor and execute the model from a command line on their local machine.

Understanding that it was crucial to address this issue, the virtual watershed protoype system was developed with a web-based scenario design tool. This tool provides a handy interface which enables for the rapid adjustment and calibration of PRMS model input variables. It also provides the tools to execute the tweaked models via the modeling web service, and visually compare outputs from differently calibrated runs. Figure 7 shows the interface developed which provides visually oriented tools for data manipulation. This interface abstracts away technical model data representations and allows users to model in intuitive ways. Figure 8 shows the output visualization after a the modeling web service has finished executing the model run.

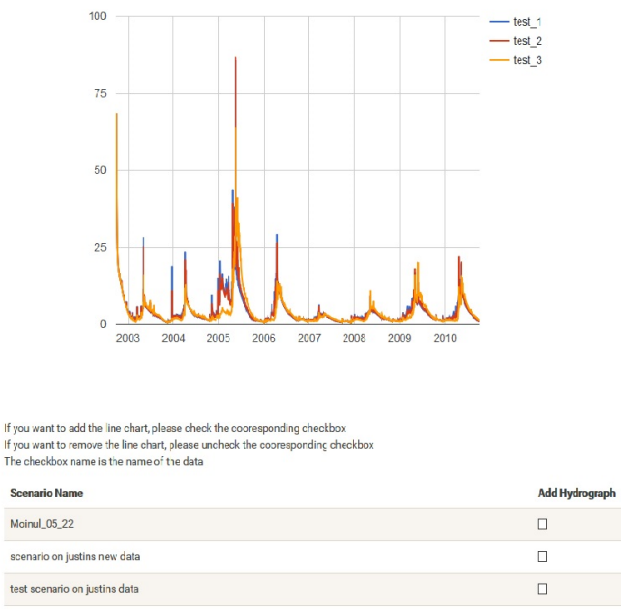


Figure 8: Comparing the results of the scenarios created using the modeling service

We developed multiple IPython [30] notebooks to demonstrate the programmatic approach to running an ensemble of models in parallel by using the modeling API by Matthew Tuner, who is a PhD Student in Cognitive and Information Sciences at University of California, Merced. Figure 9, shows an important step of modeling with the PRMS model. By using the modeling API, an user can execute multiple models in parallel, programmatically. Users can use server resources to run a model many, many times without concerning themselves about limited time and CPU resources.

```
import itertools
fig, ax = plt.subplots()

cax = ax.matshow(nash_sutcliffe_mat, cmap='viridis')
tix = [1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0]
plt.xticks(range(10), tix)
plt.yticks(range(10), tix)


ax.xaxis.set_ticks_position('bottom')
plt.ylabel('jh_coef_hru factor')
plt.xlabel('rad_trncf factor')

for i, j in itertools.product(range(10), range(10)):
    plt.text(j, i, "%.3f" % nash_sutcliffe_mat[i, j],
             horizontalalignment="center",
             color="w" if nash_sutcliffe_mat[i, j] < .54 else "k")

plt.title('Nash-Sutcliffe Matrix')
plt.grid(b=False)
cbar = fig.colorbar(cax)
```

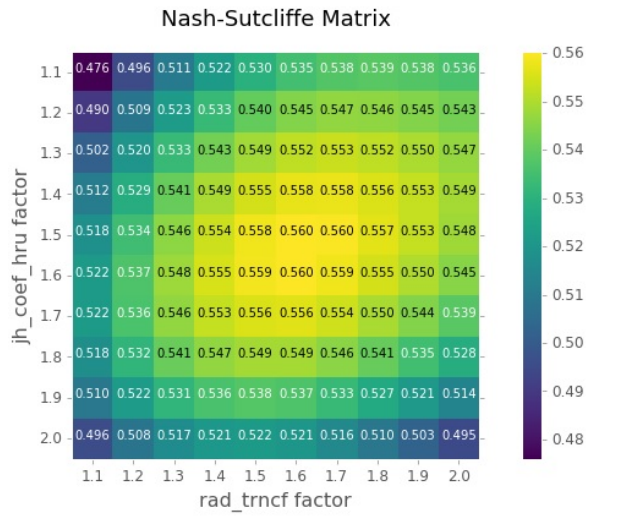


Figure 9: Example of tuning a hundred model through an IPython notebook with different parameters

5 Comparison with Related Tools

A feature based comparison is made between similar software tools focused on hydrologic and environmental modeling in Table 1. The tools discussed in Section 2 are CSDMS [8], Hydroshare [9] and MAAS [10]. Though each of these tools were designed and developed with different goals in mind, all of them are a demonstration of work to assist environmental modeling in general. Each of these tools has their pros and cons from different users perspectives.

CSDMS is a community driver system where modelers can submit their models to CSDMS by implementing model interfaces provided by them. The model gets evaluated and added into the system by the CSDMS committee. CSDMS uses a language interoperability tool called Babel to handle language heterogeneity. From the user's perspective, CSDMS provides a web and desktop based client where users access models and run them using different configurations from within the client. On the server side, CSDMS maintains an HPC cluster where they have all the models and relevant dependencies installed.

Hydroshare [9], on the other hand, is a project started to accommodate data sharing and modeling for hydrologists. The platform is in active development and frequently adds new features. The current version of Hydroshare is more concentrated towards data sharing and discovery for hydrology researchers [31]. Hydroshare also provides programmatic access to its REST API which makes it a good candidate for a data discovery backend for any other similar system.

The MaaS [10] is the most similar project to the Virtual Watershed modeling tool. The main advantage our work has over MaaS is resource management. Dr. Li et al. proposed MaaS, but they use Virtual Machine techniques in their framework. In our system, Docker containers are used, which require fewer resources. Also, we implemented APIs to check and stop a docker container based on model execution status. Existing container orchestration tools, like Docker Swarm, are limited in thier ability to manage containers in this way. Though the approach and implementation of this work differ in various ways, the end goal is similar to what we are trying to achieve. MaaS introduced models as services by creating an infrastructure using cloud platforms. The framework provides users with a web application to submit jobs. The backend that takes care of on-demand provisioning of virtual machines that containing model execution environment setup. MaaS achieves model registration by encapsulating a model as a virtual machine image in an image hub. It provides an FTP based database backend to store the model results.

6 Conclusion and Future Work

A hydrologic model execution web service platform named VWS is described in this paper. The key idea of the proposed platform is to expose model relevant services through python wrapper. This wrapper enables representation of model resources with a common data format (e.g., NetCDF), increasing inseparability. Also, it requires NetCDF format resources, which simplifies model execution on a server node. A user can login once and access all services of the VWS with the authentication/authorization component. Each VWS Component is wrapped in a docker container, which requires less resources than a traditional Virtual Machine. Docker container APIs, including the container termination API, are also implemented to autonomously manage the dynamic resource needs of this system.

This software has significant room to be expanded upon with furutre work. For example, the system currently only provides the option to integrate with GSToRE data backend. This can be improved through implementation of a generic data backend which enables integration with other existing data providers like DataONE, Hydroshare, and CUHASI. This, in turn, could provide better access to data for modeling automation. Developing a pricing component for service usage would be considered for future development. This aspect can help system manager to sustainably

Table 1: Feature Comparison with Related Hydrologic and Environmental Modeling tools

Feature \ Tool	CSDMS	HYDRO-SHARE	MAAS	VW-MODEL
Open source	Yes	Yes	No	Yes
Provides REST API	No	Yes	No	Yes
Provides interface for model implementation	Yes	No	No	Yes
Allows model coupling	Yes	No	No	No
Allows model registration	Yes	No	Yes	Yes
Data storage	No	Yes	Yes	Yes
Data sharing and discovery	No	Yes	No	No

maintain the server and hire software developers to implement other useful features.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We would like to thank Matthew Turner for his sample python scripts described in Section 4 and all the reviewers for their suggestions.

This material is based upon work supported by the National Science Foundation under grant numbers IIA-1301726 and IIA-1329469.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Nmepscor.org. The western consortium for water analysis, visualization and exploration — new mexico epscor., 2018. [Accessed on 10 June 2018].

[2] Md Moinul Hossain, Rui Wu, Jose T Painumkal, Mohamed Kettouch, Cristina Luca, Sergiu M Dascalu, and Frederick C Harris. Web-service framework for environmental models. In *Internet Technologies and Applications (ITA), 2017*, pages 104–109. IEEE, 2017.

[3] Russ Rew and Glenn Davis. Netcdf: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82, 1990.

[4] Danny Marks, James Domingo, Dave Susong, Tim Link, and David Garen. A spatially distributed energy balance snowmelt model for application in mountain basins. *Hydrological processes*, 13(12-13):1935–1959, 1999.

[5] Steven L Markstrom, Robert S Regan, Lauren E Hay, Roland J Viger, Richard M Webb, Robert A Payn, and Jacob H LaFontaine. Prms-iv, the precipitation-runoff modeling system, version 4. Technical report, US Geological Survey, 2015.

[6] Chao Chen, Ajay Kalra, and Sajjad Ahmad. Hydrologic responses to climate change using downscaled gcm data on a watershed scale. *Journal of Water and Climate Change*, 2018.

[7] Chao Chen, Lynn Fenstermaker, Haroon Stephen, and Sajjad Ahmad. Distributed hydrological modeling for a snow dominant watershed using a precipitation and runoff modeling system. In *World Environmental and Water Resources Congress 2015*, pages 2527–2536, 2015.

[8] Scott D Peckham, Eric WH Hutton, and Boyana Norris. A component-based approach to integrated modeling in the geosciences: The design of csdms. *Computers & Geosciences*, 53:3–12, 2013.

[9] David G Tarboton, R Idaszak, JS Horsburgh, D Ames, JL Goodall, LE Band, V Merwade, A Couch, J Arrigo, RP Hooper, et al. Hydroshare: an online, collaborative environment for the sharing of hydrologic data and models. In *AGU Fall Meeting Abstracts*, 2013.

[10] Zhenlong Li, Chaowei Yang, Qunying Huang, Kai Liu, Min Sun, and Jizhe Xia. Building model as a service to support geosciences. *Computers, Environment and Urban Systems*, 61:141–152, 2017.

[11] Michael P McGuire and Martin C Roberge. The design of a collaborative social network for watershed science. In *GeoInformatics in Resource Management and Sustainable Ecosystem*, pages 95–106. Springer, 2015.

[12] Eric Fritzinger, Sergiu M Dascalu, Daniel P Ames, Karl Benedict, Ivan Gibbs, Michael J McMahon, and Frederick C Harris. *The Demeter framework for model and data interoperability*. PhD thesis, International Environmental Modelling and Software Society (iEMSs), 2012.

[13] Jeffrey D Walker and Steven C Chapra. A client-side web application for interactive environmental simulation modeling. *Environmental Modelling & Software*, 55:49–60, 2014.

[14] Jon Wheeler. Extending data curation service models for academic library and institutional repositories. *Association of College and Research Libraries*, 2017.

[15] Thomas Erl. Service-oriented architecture (soa): concepts, technology, and design. Prentice Hall, 2005.

[16] Roy T Fielding. *Architectural styles and the design of network-based software architectures*, volume 7. University of California, Irvine Doctoral dissertation, 2000.

[17] Alex Rodriguez. Restful web services: The basics. *IBM developerWorks*, 33, 2008.

[18] Leonard Richardson and Sam Ruby. *RESTful web services*. O'Reilly Media, Inc., 2008.

[19] Peter Buxmann, Thomas Hess, and Sonja Lehmann. Software as a service. *Wirtschaftsinformatik*, 50(6):500–503, 2008.

[20] James Lewis and Martin Fowler. Microservices: a definition of this new architectural term. *MartinFowler. com*, 25, 2014.

[21] Sam Newman. *Building Microservices*. O'Reilly Media, Inc., 2015.

[22] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.

[23] Md Moinul Hossain. *A Software Environment for Watershed Modeling*. PhD thesis, University of Nevada, Reno, 2016.

[24] Rui Wu. Environment for Large Data Processing and Visualization Using MongoDB. Master's thesis, University of Nevada, Reno, 2015.

[25] Miguel Grinberg. *Flask web development: developing web applications with python*. O'Reilly Media, Inc., 2018.

[26] Rick Copeland. *Essential sqlalchemy*. " O'Reilly Media, Inc.", 2008.

[27] Bruce Momjian. *PostgreSQL: introduction and concepts*, volume 192. Addison-Wesley New York, 2001.

[28] Ask Solem. Celery: Distributed task queue., 2018. [Accessedon 10 June 2018].

[29] Docker. Docker hub, 2018. [Accessed on 10 June 2018].

[30] Fernando Pérez and Brian E Granger. Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, 9(3), 2007.

[31] Daniel P Ames, Jeffery S Horsburgh, Yang Cao, Jiří Kadlec, Timothy Whiteaker, and David Valentine. Hydrodesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. *Environmental Modelling & Software*, 37:146–156, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 391-402 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Robot Self-Detection System

Ivaylo Penev[*,1], Milena Karova[2], Mariana Todorova[1], Danislav Zhelyazkov[1]

[1]*Technical University of Varna, Department of Computer Science and Engineering, 9010 Varna, Bulgaria*

[2]*Technical University of Varna, Department of Automation, 9010 Varna, Bulgaria*

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 18 November, 2018
Online: 07 December, 2018

Keywords:
Robotics
Image recognition
2D coordinates
3D coordinates
Arduino

ABSTRACT

The paper presents design and implementation of a mobile robot, located in an accommodation. As opposed to other known solutions, the presented one is entirely based on standard, cheap and accessible devices and tools. An algorithm for transformation of the 2D coordinates of the robot into 3D coordinates is described. The design and implementation of the system are presented. Finally, experimental results with different devices are shown.

1. Introduction

This paper is an extension of work, originally presented in [1].

Robot orientation is a field of scientific and practical interest for many years. Different methods and approaches for different instances of the robot orientation problem are proposed and studied, e.g. [2-9]. Most often communication technologies for robot orientation and path planning are used, for example RFID [10-12].

On the basis of the literature review several main challenges, concerning robot orientation, arise.

How to recognize the robot, using images of the robot and the environment;

This is a problem of image analysis and computer vision areas. Different methods for image cutting and segmentation are known, e.g. [13, 14]. The algoritms for robot localization use lazers, sonar senzors and stereo vision systems, e.g. [15, 16]. Although these solutions show good results, it is not always possible to provide the robot with lazers, sonars or vision system.

How to convert the 2D into 3D coordinates;

There known algorithms for transformation of coordinates from 2D to 3D plane rely on lanzer range finders or vision systems, e.g. [17-19].

How to design the applications, concerning their usage and scalability.

Image processing of robot and environment usually require many resources of the robot device (CPU and RAM). Most of the known applications rely on groups of robots or mobile agents, e.g. [20-23]. Of course, it is not always possible to provide groups of robot devices for experiments.

As a final conclusion of the problems and the known solutions, described above, we could summarize, that there are not complete and working solutions of the robot orientation problem, based on ordinary devices with limited resources.

The current work presents a robot orientation system, using a cheap Arduino based robot, supplied with LED sensors, and an ordinary mobile device. The paper layout is concentrated on three main points:

- Robot recognition from the image, shot by the device's camera;
- Converting the 2D coordinates to 3D;
- Implementation of the system by wide spread hardware and programming tools.

The purpose of the presented approach is to provide methodology for applying the robot orientation problem, using standard, cheap, wide spread devices.

[*]Corresponding Author: Ivaylo Penev, Email: ivailo.penev@tu-varna.bg

2. Methodology

2.1. System overview

The "Selfie robot" consists of a robot and a mobile device with a camera. The camera takes a picture of the robot and sends it to the robot. The robot's control unit computes its self-coordinates as well as the camera's coordinates. Afterwards it builds a path for moving to the center of the frame.

The robot is supplied with LED sensors. The system is implemented by standard Arduino architecture and OpenCV library.

2.2. Robot recognition

The Robot recognition algorithm consists of several separate steps together solving the common problem.

1) Robot marking

Simple yet efficient approach should be using light-emitting diodes (LEDs). LED is an electronic diode, which converts electricity to light – this effect is called electroluminescence. When an appropriate voltage is applied to it an energy is being disposed in the form of photons. The color of the light is defined by the semiconductor's energy gap.

Simplified LEDs are colorful light sources which makes them easy for recognition in environments where light is reduced. Marking the robot with such diodes could make it look unique regardless of the daylight. The solution should work regardless of the environment's brightness.

2) Camera properties

The previous pictures are taken by a camera, which properties are set to default – most commonly to auto-adjustment values, which ensures better quality of the picture. The current project does not depend on high end quality. All that is needed is a recognition of the robot.

The following settings deal with light perception:

a) ISO

This parameter measures the the camera sensitivity regarding the light.

b) Shutter speed

The shutter speed measures the time period necessary to fire the camera.

c) Aperture

Aperture (also known as *f-number*) presents a port into the lens, thorugh which light moves into the body of the camera.

d) Exposure value

The value of this parameter is determined by the aperture and the shutter speed in such way ensuring that combinations giving the same exposure have the same exposure value (1).

$$exposure\ value = \log_2\left(\frac{aperture^2}{shutter_speed}\right) \quad (1)$$

Practically the exposure value indicates how much a picture is illuminated or occulted.

Given these definitions a simple approach comes up for reducing the unnecessary light from the taken pictures. They could be darkened by reducing the ISO and Exposure values.

Figure 1 and Figure 2 present photos, taken with ISO set to 50 and exposure value set to -2. An improvement can be easily noticed. By setting the ISO to a low value the camera sensor's sensitivity was reduced. By decreasing the exposure value less light enters the camera. The effect of the reflected light from solid objects was highly filtered while the light from the LEDs has not changed at all.

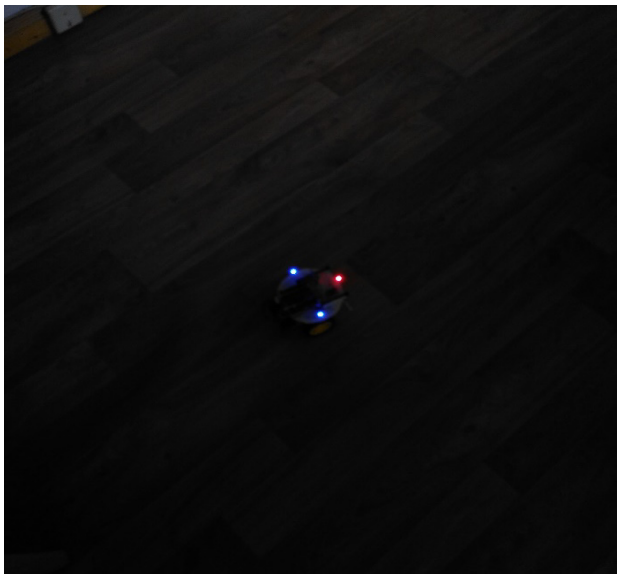


Figure 1. Dark room; reduced light sensitivity

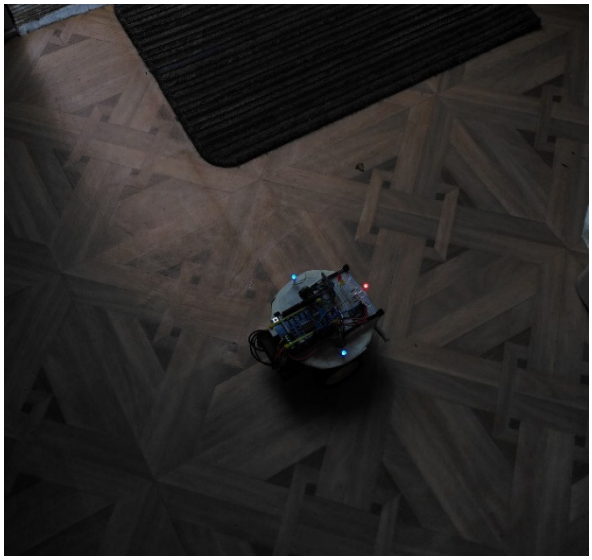


Figure 2. Light room; reduced light sensitivity

The LEDs' light now can be easily extracted even if the robot is in a brighter environment. However they are seen as little blobs and the light reflected by highly specular objects could be still present as noise, thus presupposing algorithm confusion and failure.

3) Camera focus

Most cameras support different focus modes, which helps the user adjust the focus:

- Auto-focus: automatically finds the best focus range, where most of the objects are sharp;
- Infinity: the focus is set to the farthest range leaving close objects dull;
- Macro: the focus is set to the closest range making nearby objects look sharp;
- Manual: leaves focus range to be adjusted manually be the user.

Figure 3 is taken with ISO 50, EV -2 and Macro focus in a bright environment. It appears, that reducing the camera focus not only blurs / removes noise from strong reflections, but also improves the LEDs light enlarging and saturating their blobs.



Figure 3. Lighter room; defocused with reduced light sensitivity

4) Image analysis

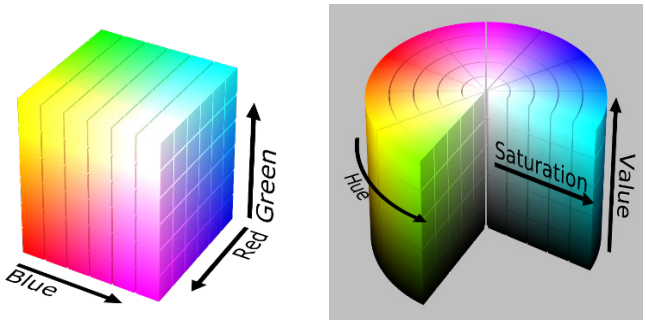
So marking the robot with LEDs and adjusting the camera settings ensures more or less the same picture format with clearly expressed form of the robot regardless of the environment's illumination. The taken pictures are analyzed with a computer vision algorithm. The taken approach relies on the following two color ranges:

a) RGB

Red-green-blue color model is an additive color model consisting of red, green and blue lights which combined give a wide range of different colors. This model is the base of the color space and is considered as the most identical and the easiest to understand by the human eye. Identical values of the three lights give a shade of the gray color. Keeping up the lights (at least one) gives a bright color while if all three of them are low will result in a dark color, close to the black one. In electronic devices a color in that space is represented by 3 Bytes – a Byte for the value of each light. Could be graphically represented by a cube (Figure 4a).

b) HSV

Hue-saturation-value is another range based on the RGB model. It is considered to be more convenient for usage in computer graphics and image editing. Hue represents the color intensity, saturation – the color completeness and the value is often call the brightness of the color. This color space is one of the most common cylindrical representation of the RGB model. Like the RGB color space a color of this space can be also stored in 3 Bytes – one for each channel (Figure 4b).



a) RGB color model b) HSV color model

Figure 4. Color model

The goal is to extract the **blue** LEDs for example. First an image representing the blue channel from the RGB color space is obtained (Figure 5).



Figure 5. RGB Blue channel

The image is greyscale because it has only a single channel. The blue blobs can be easily noticed, because they are colored in high grey value. But so does the door's threshold. In fact every whiter object from the original image will have a high grey value in the blue channel image, because the white color in the RGB color space consists of high values in all of the three channels – red, green, blue.

The difference between the blue LEDs and the door's threshold is obvious – LEDs are more colorful in the input image. A good representation of the color intensities is the *saturation* channel form the HSV color variant of the input image (Figure 6).

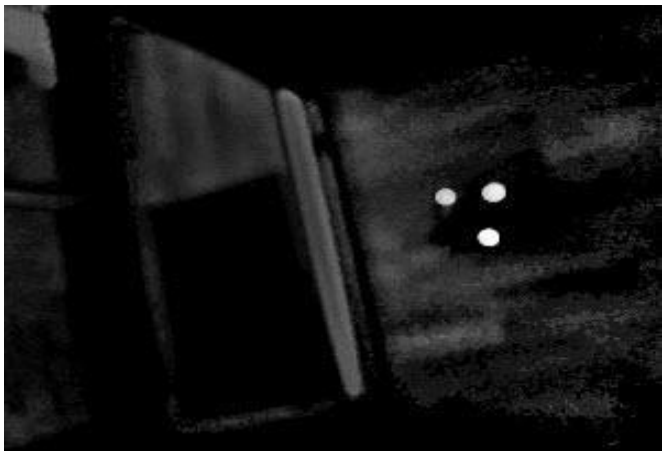


Figure 6. HSV Saturation channel

The difference is increased. The values of the LEDs are much higher than the one of the door's threshold. The saturation image could be a perfect mask for ignoring white objects in all of the three channels from the RGB color space.

Next a masking of the blue channel is performed with the saturation image – applying a logical AND between the images' binary data (Figure 7).

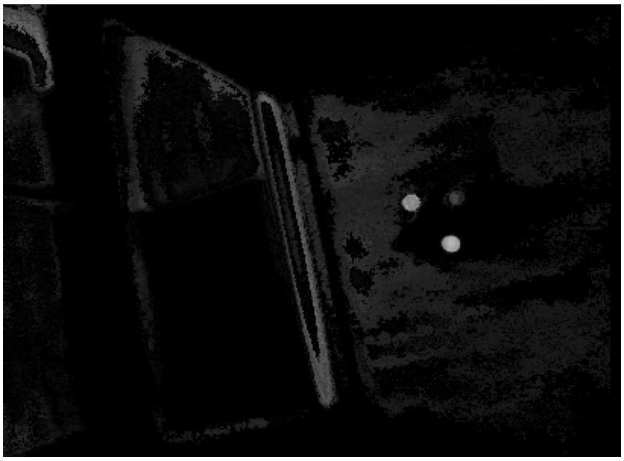


Figure 7. Merged blue and saturation (blue_sat) channels

c) Binary image

Binary images are simple images whose pixels can have only two possible states – 0/1 or black/white. In most cases white pixels are considered as part of an object or of the foreground of the image and black – as the background. In computer vision binary images are mostly used as a result of different processing operations such as segmentation, thresholding, and dithering.

d) Image segmentation

Image segmentation is a technique to divide a digital image into multiple zones. The purpose is to detect objects and boundaries within the image.

e) Thresholding

Thresholding is one of the simplest image segmentation methods. It basically separate a greyscale image into a background and foreground areas as a resulting binary image. The algorithm is simple – if a pixel value is higher than a given constant (called threshold), it is classified as a part of the foreground regions and its value is set to 1 in the binary image. Otherwise the pixel is considered as a background one and a 0 is assigned to its value. The most challenging task while using this method is determination of a correct threshold value. Often a technique called Otsu's method is used to solve this task. It calculates optimal threshold value by finding the minimal intra class variance between two class of points – one with lower values and one with higher values.

These image processing techniques are used by the LEDs recognition algorithm. After masking white regions in the blue image, a binarization is required. After the filtering a lot of dark regions have occurred. For certain LEDs' pixel values are highest. Histogram analysis must be done only in the upper half of the histogram.

This is the extracted upper half of the merged image. Almost every time the LEDs' pixel values are represented by the first highest peak. Otsu's method is used over this part of the histogram. The calculated threshold is colored in green (Figure 8).

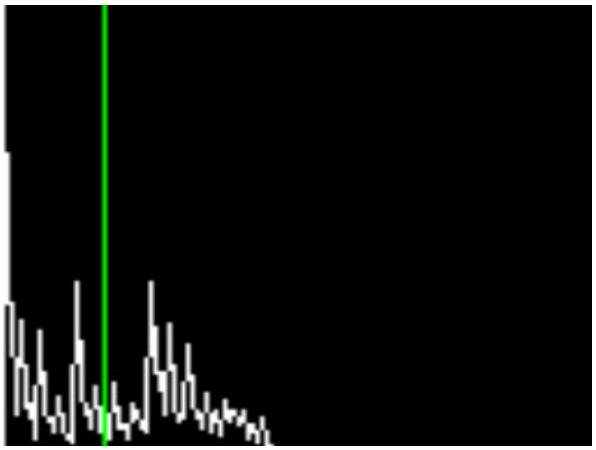


Figure 8. Blue_sat histogram

This binary image is a result of thresholding and consequential dilation. With the use of connected-component labeling technique regions could be easily extracted and their centroids could be computed by the average point formula. These centroids represent the {x, y} coordinates of the blue LEDs (Figure 9).

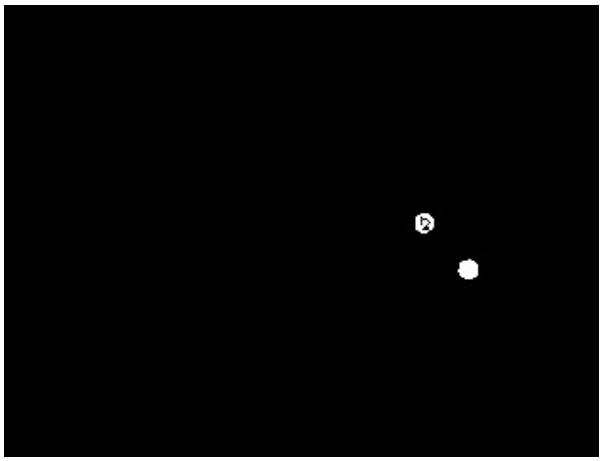


Figure 9. Blue_sat binary image

The same algorithm could be applied for the other two RGB channels. Since there are no green LEDs results will be shown only for the red channel (Figure 10, Figure 11).



Figure 10. RGB red channel



Figure 11. Merged red and saturation (red_sat) channels

At the end {x, y} coordinates of the red and blue LEDs are present which makes the task of recognizing the robot completed.

2.3. Transformation of 2D coordinates to 3D

The camera of the device shoots the robot's sensors. The coordinates of the robot are calculated and sent to the robot using the Bluetooth interface. This process is presented at Figure 12.

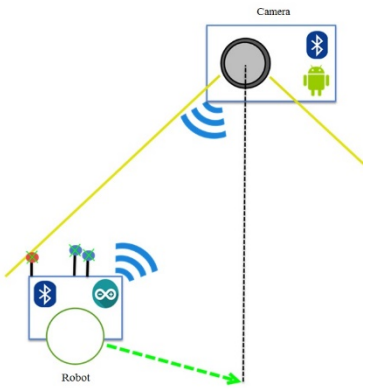


Figure 12. Sending coordinates to the robot

The robot moving to the goal consists of turning and moving forward. Turning uses the angle between the direction of the robot and the goal direction. The LEDs coordinates, extracted from the image by the recognition algorithm, put in a Cartesian coordinate system result in the following model (Figure 13):

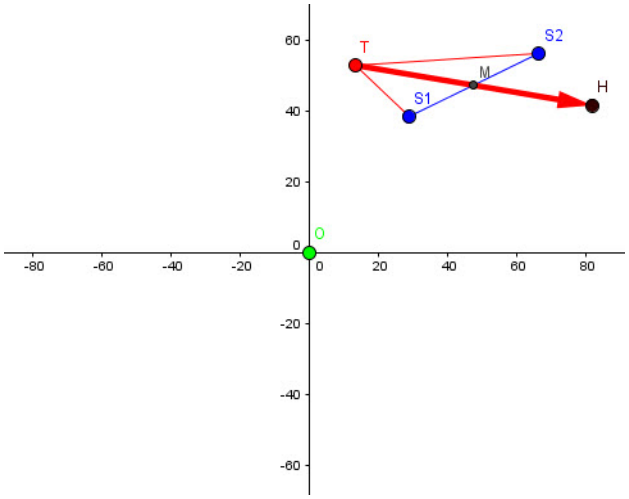


Figure 13. Extracted LEDs coordinates in Cartesian coordinate system

The model is an approximate example. For increasing the detail in the interest area future models will be limited only to the first quadrant.

$\overline{TH}$

The environment in which the robot and the camera exist is three dimensional. Two dimensional object representations in such 3D world exist in so called planes. The camera takes 2D images which belongs to a plane determined by the camera. The robot on the other hand is a 3D object. Its representation in the 2D image is a projection of itself in that plane. Robot's LEDs are attached in a way that keeps them on equal distance from the floor. It could be assumed that they are laying in a 2D plane parallel of the ground plane. What appears is that the extracted coordinates of the LEDs are their projections on the phone's plane, moreover the center of the image is a projection of the robot's target.

The following task arises: determine the robot's position and orientation against its target by the given projections' coordinates. After little mathematical analysis it appears that these coordinates are not enough for solving that given task. But that is not all. The robot knows the position of its LED sensors (Figure 14).

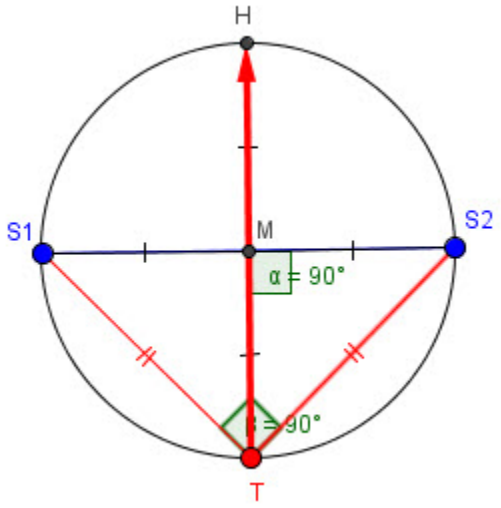


Figure 14. Robot representation in its plane

The symbols in the diagram have the following meaning:

- T – tail point, representing the red LED with the extracted coordinates,
- S1, S2 – side points, representing the blue LEDs with the extracted coordinates,
- – the center of the image/coordinate system, representing the goal's position,
- M – midpoint of the S1S2 segment, representing the center of the robot around which the robot should turn
- H – the T's mirror point against M, representing the head of the robot,
- $\overline{TH}$ – a vector indicating robot's size and the direction it is looking at.

This model is representation of the robot in its plane. It knows where are its head, center and tail, what size is its diameter and that the LEDs form an isosceles right triangle. All that is left to be

found is the target's coordinate against the robot in that plane – the O point. After that the value of the turn angle is defined by the points H, M, O. The distance to O is defined by (2):

$$distance = k.robot_diameter \quad (2)$$

where $k = \frac{\overline{MO}}{\overline{HT}}$.

Since the robot knows its physical diameter and the distance depends on a proportional variable the robot's plane doesn't need to be the same size – just the same proportions.

2.4. Analytic Geometry for Transformation

The technique for accomplishing the task is simple and relies on basic laws from the Analytic geometry. Analytic geometry studies geometry using coordinate system – mostly the Cartesian one to deal with equations for planes, lines, points and shapes in both two and three dimensions. It defines, represents and operates with geometrical shapes using numerical information.

The current algorithm operates only with 2D coordinates {x,y} and the following shapes are being used:

Point

Represented by one pair of {x, y} coordinates which defines its position on the coordinate system.

Angle

Represented by single real value in radians or in degrees in the intervals respectively [0, 2π) or (-π, π] and [0, 360) or (-180, 180].

Line

Represents an unlimited straight ray. It is defined by (2):

$$y = K.x + C \quad (2)$$

, where

K – the tangent of the angle between $Ox^{\rightarrow}$ and the line; it has a constant value,

C – the offset between the intersection of $Ox^{\rightarrow}$ and the line and the center point, it has a constant value.

Laws

The following laws are used by the algorithm:

Line from two points p1 and p2 (3)(4)

$$K = \frac{y_2 - y_1}{x_2 - x_1} \quad (3)$$

$$C = y_1 - K.x_1 \quad (4)$$

Line parallel to line l_1 passing through point p_1 (5)(6)

$$K = K_1 \quad (5)$$

$$C = y_1 - K.x_1 \quad (6)$$

Intersection point of two lines l_1 and l_2 (7)(8)

$$x = \frac{C_2 - C_1}{K_1 - K_2} \quad (7)$$

$$y = K_1.x + C_1 \quad (8)$$

Distance between two points p1 and p2 given by the Pythagorean theorem (9)

$$distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (9)$$

Angle α from three points p1, p2 and p3 (10)

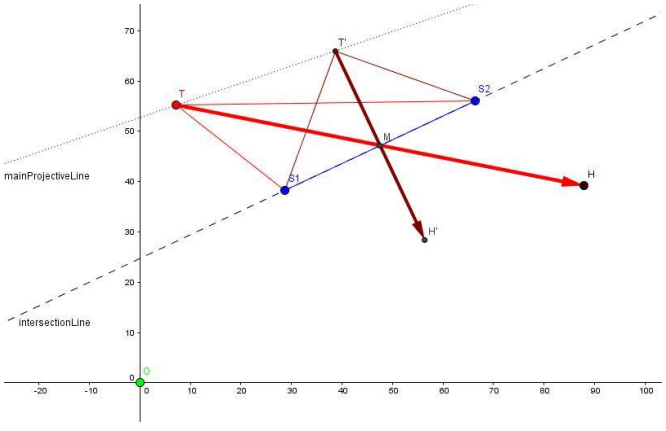
$$\alpha = \alpha tan2(y_1 - y_2, x_1 - x_2) - \alpha tan2(y_3 - y_2, x_3 - x_2) \quad (10)$$

where

$$\alpha tan2(y,x) = \begin{cases} \arctan\left(\frac{y}{x}\right), if\ x > y \\ \frac{\pi}{2} - \arctan\left(\frac{x}{y}\right), if\ y > 0 \\ -\frac{\pi}{2} - \arctan\left(\frac{x}{y}\right), if\ y < 0 \\ \arctan\left(\frac{y}{x}\right) \pm \pi, if\ x > 0 \\ undefined, if\ x = 0\ and\ y = 0 \end{cases}$$

2.5. Transformation Algorithm Definition

There are two planes – the phone's one and the robot's one, several laws, extracted points and real proportions. The two planes have to intersect somewhere in the space. If not that means they are parallel – the camera is parallel to the ground and the searched target's position is in fact the center of camera image. But in most cases it is not and the two planes intersect. If the actual intersection line is found, distance from the camera to the robot could be computed. But since the result could be proportional that operation is unnecessary and the intersection line could be any line from the two planes. Of course the shapes need to keep original robot proportions. It is assumed that the intersection line is the one that passes through the two blue points giving the following model as a result (Figure 15).



Intersected planes with an example projection line

The dark color has shapes in the robot's plane while brighter ones – in the image plane; a shape's projection has the same name with suffix '.

T', *H'* – projections of *T* and *H*, showing how would *T* and *H* coordinates be if the image's plane was parallel to the robot's one.

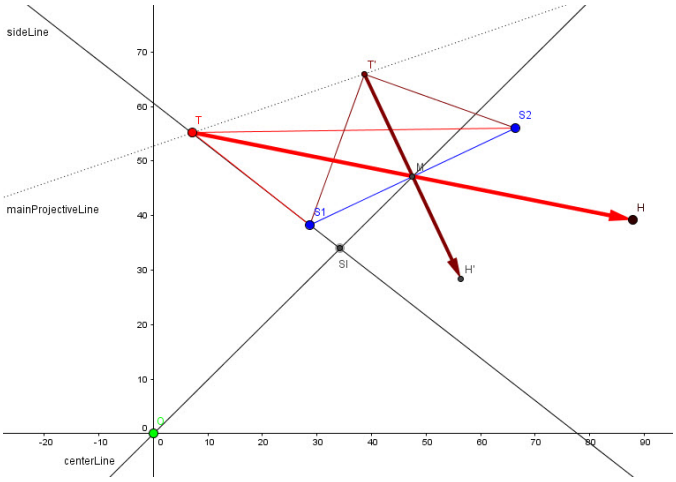
As it can been seen on Fig. 4 *S1*, *S2* and *T'* also define an isosceles right triangle in 2D – the proportions are kept.

A line could be defined from either *T* and *T'* or *H* and *H'* and could be called "main projective line". All newly defined projective lines have to be parallel to this main one.

Let's define the following shapes (Figure 16):

- Line from *O* and *M* called "center line";
- Line from *T* and either called "side line" that can intersect the "center line";

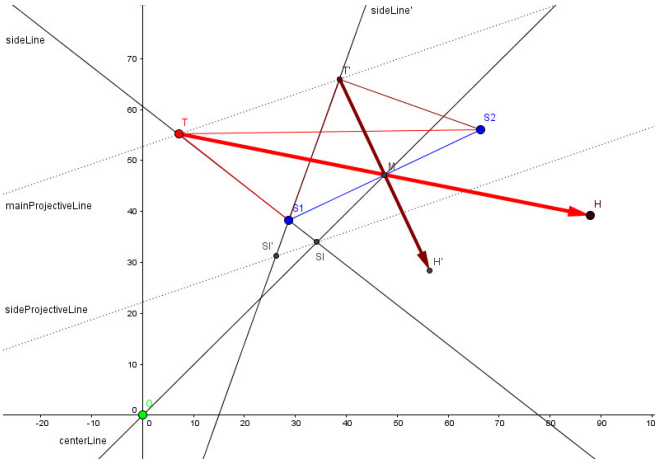
Intersection point SI from the center and side lines.



The "side line" is defined using *S1* which in this case is correct since the side line is not parallel to the center one. But there is a possibility that the line from *T* and *S1* may not intersect the center line. If such case occurs the side line should be defined using *S2*. For certain *T*, *S1*, *S2* form a triangle so if *S1* is not appropriate for defining the side line, *S2* will definitely be.

Next step consists of defining the following shapes (Figure 17):

- Side line's projection – line from *T* and *S1* (or *S2* if *S1* was not relevant during the previous step);
- Projection point of *SI* on the side line's projection.

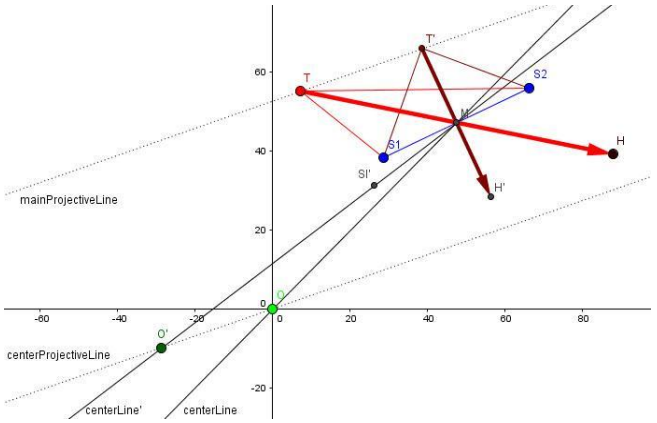


Projecting the side intersection line

Defining a projection of a point on a line consists of two steps. Firstly a projective line must be defined. That line passes through the projected point and is parallel to the main projective line. The projection point is considered the intersection point of the projective line and the line that it is supposed to lie on. In this case the projected point is *SI*, the projective line is *sideProjectiveLine* and the projection point is *SI'* that lies on *sideLine'*.

The final algorithm step consists of defining the following shapes (Figure 18):

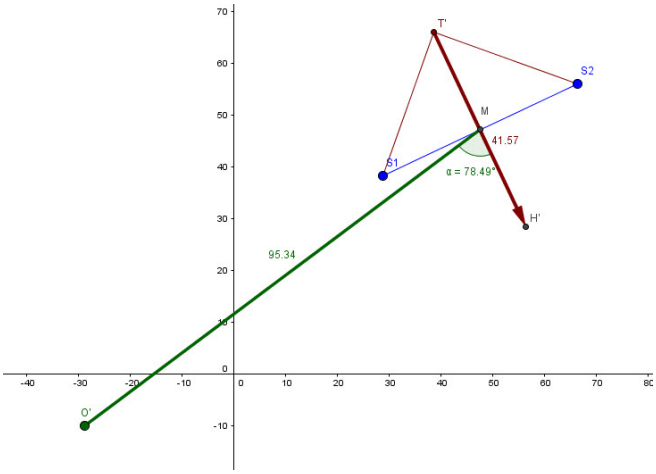
- Define a line from *M* and *SI'*; since *SI'* is a projection of *SI* along with the fact that *SI* and *M* lay on the center line, it could be stated that this new line is the projection of the center line;
- Define the projection of *O* on the center line's projection.



O' is the projection of *O* which is defined the same way *SI'* was. But this time the projective line is *centerProjectiveLine* and the second line for intersection is *centerLine'*.

O' now represents the location of the target in the robot's plane.

The result projection is shown on Figure 19.



Given that point the angle *H'MO'* could be easily computed using the declared law.

The lengths of the segments *MO'* and *H'T'* are also known which enables the evaluation of (1):

$$k = \frac{\overline{MO'}}{\overline{H'T'}} = \frac{95.34}{41.57}$$

3. System requirements and specification

3.1. Mobile robot

The robot used in this project is based on the Arduino platform. It is an open-source microcontroller-based kit for building digital devices and interactive objects that could interact with the environment with sensors and actuators. Arduino board provides set of analog and digital I/O pins which enable connectivity with external physical devices. Programs for Arduino are written in either C, C++ or Processing.

3.2. Arduino board

Robot's controller board is an Arduino UNO R3, which is a basic microcontroller board, suitable for simple startup projects. It supports Universal Serial Bus (USB) with the help of which a serial communication could be established with PC and other devices and could also supply power to the board. Usually programs are built and compiled on a PC and further transferred to the board using the USB. It has also a 2.1mm center-positive power jack for external power supplies.

The board, used for assembling the robot, has the following parameters:

- Microcontroller: ATmega328
- CPU: 8-bit AVR
- Clock: 16MHz
- Memory of 23KB flash, 2KB SRAM and 1KB EEPROM
- 14 I/O pins
- 6 Analog Input pins
- Physical parameters: 68.6x53.4mm , 25g weight

3.3. Robot body

For basic movements a simple 2 Wheel Drive (2WD) structure is used. This is a DIY 2WD double leveled plastic chassis. It consists of 2 wheels with separate gear motors, full-degree rotating caster wheel as a third strong point and two decks. The wheels are positioned almost in the center of the structure. Their rotation in opposite directions will result in the whole structure spinning over its center. The two decks are ideal for separating the mechanics from the electronics.

3.4. Motor shield

The motor shield of the robot provides simple motor control. It provides a control of a motor's speed, direction and braking along with sensing of the absorbed current. This board is stacked over the Arduino board.

3.5. Bluetooth module

The robot is supplied with additional Bluetooth communication module.

3.6. Magnetometer

HMC5883L is a Triple-axis Magnetometer (Compass) board. With the help of this sensor the robot can determine the direction it is facing. Furthermore using it the turning accuracy could be increased.

3.7. Camera

The project definition assumes that the robot establishes a communication with a remote device with a camera. It has one communication module – Bluetooth, therefore the remote device is required to have not only a camera but also Bluetooth communication technology.rduino board.

4. System Design

4.1. General Design

Figure 20 gives an overview of the system's design.

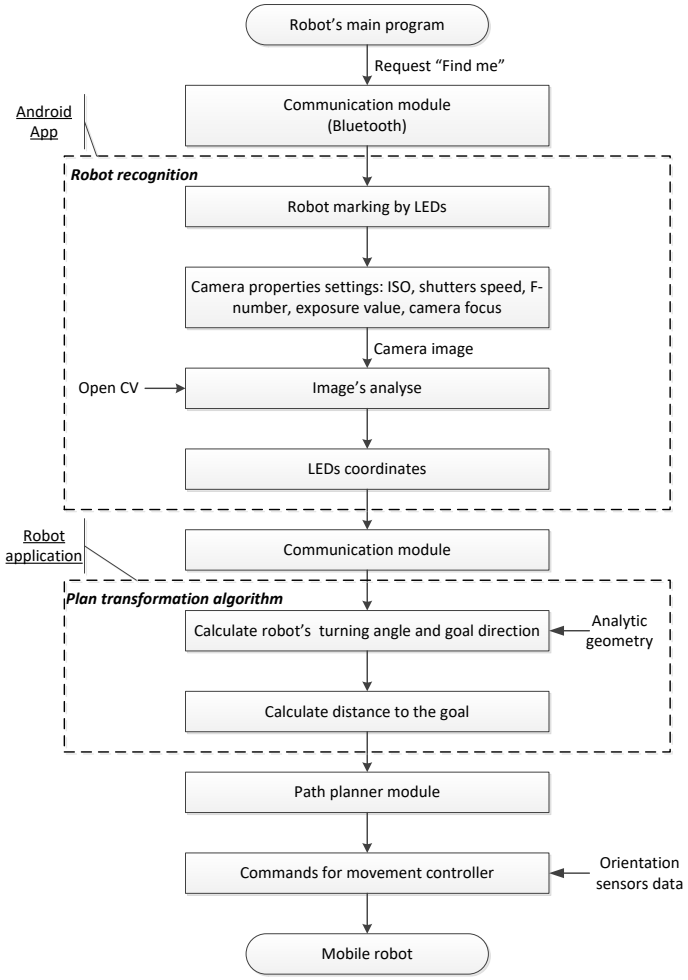


Figure 20. Flowchart of the system's general design

4.2. Robot Software Design

The software for the robot is build based on the Arduino API keeping identical design to the general one.

The design is similar to the robot side of the general diagram. The difference is that the software is built over Arduino core API while the other is more abstract.

Abstract2WDController and Motor classes are meant to be an API to the Arduino Motor shield. They encapsulate pins reservation along with basic Arduino methods for setting different digital and analog levels to the pins.

- Motor – class, representing single motor channel of the Motor shield. It adjust its pins modes and levels using Arduino basic

subroutines. Represents an API to the Motor module from the general diagram.

- Abstract2WDController – abstract class, encapsulating the declaration of two motor channels with the specific pin mapping for the Arduino Motor Shield. It also defines a basic interruptible motor control leaving the motors' parameters to be adjusted by the class' successors.
- Compass – an interface declaring a minimum behavior that one true compass object should implement.
- AngularController2 – an extension of the 2WDController providing forward movement and turning operations. In order to do accurate turns it uses a Compass object to determine its orientation. It has a predecessor sharing the same name and interface but not so successful. More information about them in the Implementation chapter.

The two controller classes could be considered as a representation of the Movement Controller module from the general diagram.

- ExSoftwareSerial – extends the Arduino's SoftwareSerial library providing methods for reading not-string based data types from the upcoming stream. These methods support error handling and time-out periods. This class represents the API to the robot's Communication module from the general diagram.
- TargetFinder – class implementing the main Plane transformation algorithm. It takes the {x,y} coordinates of the tail and the two side points and determines what turn has to be made in order to make the robot face the target and how far it is. This class does relates to the Path Planner module from the general diagram.
- Main – class representing the Main program. It binds the three features of the robot (communication, brain and motion control) together to complete the main goal of the project – staying in the center of the camera frame. This class delegates communication with the remote camera and controls the data flow from it through the TargetFinder to the Movement Controller.

4.3. Camera Software Design

The software written for the remote camera is built based on the Android API but keeps its core identical to the general idea. Keeping the fact that the software is more or less a mobile application it has to have highly intractable user interface in order to monitor the system. The design extends the general one providing additional monitoring support using Android UI tools.

4.4. Application Overall Design

This design shows how application's modules are bound together.

- Bluetooth package – this package contains all defined Bluetooth classes. These classes are façades of the Android's Bluetooth API and provides additional exception handling techniques and device discovery tools.
- Camera package – contains Camera classes extending the functionality of the Android's Camera API. It implements different tools for serving the application's specification. Camera preview class is present for handling each frame captured by the camera sensor. Additional Image processor is defined for analyzing images and a Custom camera class providing programmable interface to the project specific camera parameters for adjustment.
- Robot package – classes for handling robot requests. They implement features as robot communication managing, robot data translation and robot recognition.
- Graphical user interface (GUI) package – classes observing the core objects and updating the user interface when events occur
- MainActivity – manager class; creates all class hierarchy binding different modules together. It is responsible for resource acquiring and releasing such as Camera and Bluetooth hardware modules reserving, communicator and image processing threads handling on application starting and closing. It also appears as a controller to the main UI window granting classes form the GUI package an access to the UI objects.

4.5. Core Design

The design, presenting the core of the application, does not differ much from the general one.

- Bluetooth classes – façades to the Android's Bluetooth API; represent the interface to the device's Bluetooth (the Communication module of the mobile device).
- BluetoothModule – manager class; creates BluetoothConnection objects, deals with the Bluetooth hardware's settings and takes the responsibility of acquiring and releasing it along with closing all opened connections.
- BluetoothConnection – establishes bidirectional RF connection to a single remote device and is responsible for all data that flows in and out the device.
- RobotCommunicator – manager class; as the name suggests it completely implements the functions of the RobotCommunicator module from the general design. Additionally it uses the BluetoothModule to connect to the robot and a separate decoder called RobotTranslator to get a RobotRequestEvaluator object for handling the current request.
- RobotTranslator – factory for RobotRequestEvaluator objects; it takes the received request from the communicator and decides what evaluator to return.
- RobotRequestEvaluator – interface that unifies the type of all evaluators' responses – a byte array as specified by the BluetoothConnection's method "sendData".
- CustomCamera – façade based on the Android's Camera API providing interface to the Camera module and access to the following adjustable camera parameters (focus mode, ISO, Auto-exposure lock and exposure compensation)
- CustomCameraView – camera preview; implements frame handling that submits each new camera frame to the CameraImageProcessor and to a UI preview surface.
- CameraImageProcessor – singleton; fully implements the Image Analysis algorithm using external image processing

native library – OpenCV; the result represents all extracted red and blue regions; represents the ImageProcessor module from the general design.

- RobotLEDImageRecognizer – singleton; analyzes the extracted regions from the CameraImageProcessor trying to recognize the robot. If successful, it converts the appropriate coordinates to a binary data. If the robot is not successfully detected from the current regions, modifications of the camera's exposure are made and next result is awaited. Combinations of all available low exposure parameters are being tried.

4.6. User Interface Design

21). It consists of objects called *Observers* that are listening for activity or changes in other objects called *Subjects* or *Observables*. Observer register to a Subject which hold a list with subscribed Observers. Each time a Subject do some activity or change state it notifies all its registered listeners providing them with the required information. Using this pattern constant looping (listening) in the listener classes is avoided.

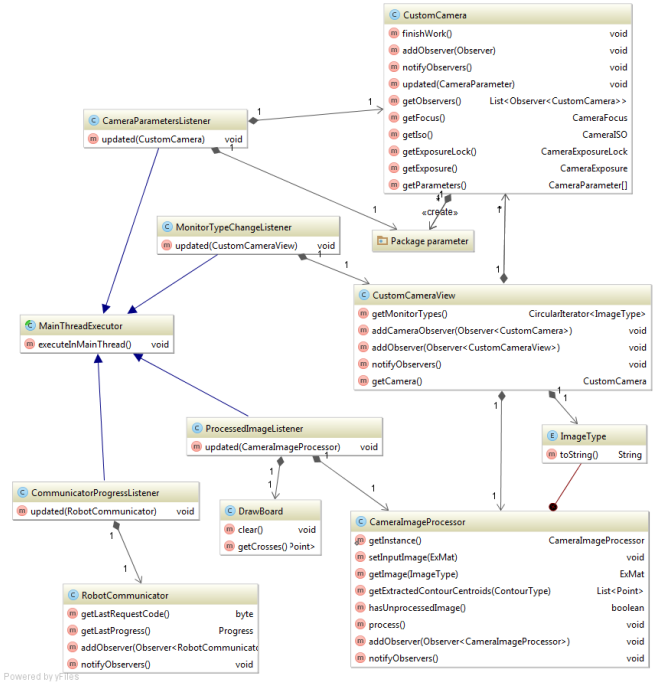


Figure 21. UML class diagram of the Android application's UI layer along with the observed core classes

- *MainThreadExecutor* – it is an abstract class providing support for executing subroutines in the main thread. Android's specification states that all interactions with user interface objects have to be done by the main thread also called *UI thread*. Also it highly recommends long-running tasks to be not executed by the main thread because it may result in UI freezing or lagging. *MainThreadExecutor* is ideal for classes that mainly interact with the UI but are being accessed by objects running in separate threads – like the *RobotCommunicator* or the *CameraImageProcessor*.
- *CommunicationProgressListener* – this class observes the *RobotCommunicator*. Each time a communication activity is being done the communicator notifies its listeners. Interaction between the communicator and its observers is being done using *Progress* typed objects. The *CommunicationProgressListener* updates an UI object each time new *Progress* object becomes present.
- *ProcessedImageListener* – an observer of the *CameraImageProcessor*. When an image has been processed the processor updates its observers. The listener retrieves the extracted regions from the picture and displays them to the screen using a *DrawBoard*. *DrawBoard* is an which represents an Android UI Surface that is positioned above any else UI object in the main UI window. It implements functionality for drawing crosses.
- *MonitorTypeChangeListener* - observes the *CustomCameraView*. By definition this custom view class handles all preview images from the Camera and sends them to both the Image processor and an UI preview class. Some additional monitoring was implemented giving the opportunity of displaying not only the original image but also intermediate processed image from the *Image analysis algorithm*. The *Image processor* provides access to such images via *ImageType* object. Clicking the view changes the type of the send image to the UI preview. The *MonitorTypeChangeListener* is notified in case of such events and reflects the displayed image type's string representation to a UI textbox.

Figure 22 demonstrates the common work of the application.

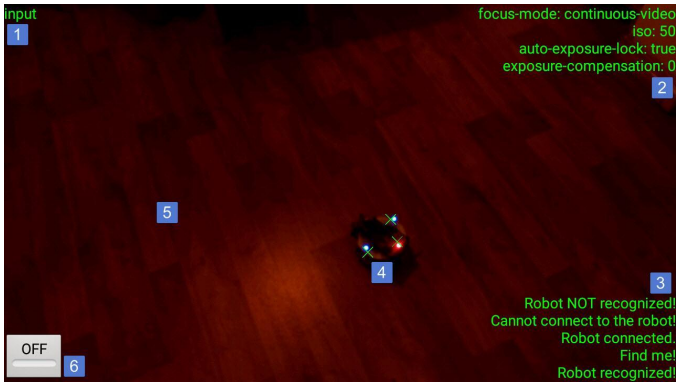


Figure 22. Screenshot of the Android application with marked UI objects

The number markers has the following meaning:

- This is a text label that is being handled by the MonitorTypeChangeListener. It displays the preview image's type.
- A multiline not-editable text area. It is modified by the CameraParametersListener each time a camera parameter is changed. Presents the current camera's settings.
- Another multiline area that keeps track of the RobotCommunicator's progress. It is being handled by the CommunicationProgressListener.

- Crosses drawn by invoking DrawBoard methods from the ProcessedImageListener.
- Touching the screen will change the preview image. The new image's type will be presented in the Label 1.
- It is a switch button that forces the ImageProcessor work even when robot communication is not present. Used for high-end monitoring but consumes more battery power.

5. Testing and Evaluation

5.1. Android Appplication

1) Robot recognition algorithm

The application was tested on three devices with quite different hardware specification. The results are summarized in Table 1.

Table 1. Camera application test results

Device model	Android	CPU	Camera parameters	Remarks
Asus ZenFone 2 Laser	API 21 v5.0 Lollipop	Qualcomm Octa-core	Focus mode ISO Exposure	Smooth performance; Fast image analysis; Robot cannot be recognized while on bright surface
LG Optimus G	API 19 v4.4 KitKat	Qualcomm Quad-core	ISO Exposure	Smooth performance; Fast image analysis; Robot cannot be recognized while on bright surface or if some noise is present.
Sony Live	API 14 v4.0 Ice Cream	Qualcomm Single-core	Exposure	Smooth performance; Fast enough image analysis; Robot cannot be recognized at all.

The following conclusions are summarized:

- The presence of several background threads doesn't badly affect overall performance on a Qualcomm CPU with less cores;
- Automatic or not-present ISO setting is not recommended;
- Support of a close focus range could ignore noise;
- The recognition depends on the supported editable camera settings. An advanced intelligent robot recognition algorith is needed to remove this dependency.

2) Application Stability

After a lot of testing and debugging it could be stated that *OpenCV*'s matrix object *Mat* is not being automatically released on garbage collection as defined in the library's documentation. Therefore it should be programmatically released before dereferenced. Otherwise a memory leak occurs which could cause system's instability.

5.2. Robot behaviour

1) Plane transformation algorithm

The plane transformation algorithm is tested with different coordinates of the LED sensors, i.e. with three input data into a Cartesian coordinate system with range [-5; -5] ÷ [5; 5].

For resolution 2 of the value of coordinates 6 values per a coordinate are possible: -5, -3, -1, 1, 3, 5. Three input points have two coordinates with $6^6 = 46656$ test cases (Table 2).

Table 2. Results from test 1

Experiments	Number of tests	Percent	Conclusion
Total	46656	100%	
Passed	36524	78.28%	
Exceptions	10132	**21.72**	
• Handled	6048	12.96%	The three points lay on one line or two of them overlap.
• Not handled	4048	8.75%	To be analyzed in future.

The tests from resolution 1 are presented in Table 3:

- each coordinate has 11 possible values in range [-5; 5];
- test cases: 116 = 1 771 561 possible combinations.

Table 3. Results From Test 2

Experiments	Number of tests	Percent	Conclusion
Total	1 771 561	100%	Conclusion
Passed	1 613 506	91.08%	
Exceptions	158 055	**8.92%**	
• Handled	84 473	4.77%	The three points lay on one line or two of them overlap.
• Not handled	73 582	4.15%	Further analysis

More combinations lead to less chance for an exception. Even a camera with low quality can take images with resolution of 1MP. This means, that the number of combinations is great and the occurrence of exceptions is practically not possible.

2) Turning accuracy

After continuous manual testing an anomaly has been discovered. The magnetometer (compass) sensor does not always gives adequate data and thus resulting in Movement controller's confusion effecting the turning accuracy. The sensor has been tested in separate project with all magnetic components (motors, battery) detached from the robot but the problem still occurs. The issue could be caused by hardware defect but another module is not available at this state.

6. Conclusions and Future Work

The results from the experimental tests prove that the system is stable and effective with different devices and resolutions of the

camera. This means, that a mobile robot, equipped with a usual camera, could be practically exploited for solving tasks as moving through dangerous spaces, searching for victims of disaster events, etc.

The future work will be direcited to implementation of intelligent software controller, using neural network for detection and recognition of the robot image. More advanced orientation sensors (accelerometer, gyroscope) will be integrated and robot's physical parameters (wheel diameter, motor RPM, robot size) will be used by the algorithm.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Karova, I. Penev, M. Todorova, and D. Zhelyazkov, "Plane Transformation Algorithm for a Robot Self-Detection", Proceedings of Computing Conference 2017, 18-20 July 2017, London, UK, ISBN (IEEE XPLORE): 978-1-5090-5443-5, ISBN (USB): 978-1-5090-5442-8, IEEE, 2017.

[2] C. Connolly, J. Burns, and R. Weiss, "Path planning using Laplace's Equation", IEEE Int. Conf. on Robotics and Automation, pp. 2101-2106, 1990.

[3] D. Lima, C. Tinoco, J. Viedman, and G. Oliveira, "Coordination, Synchronization and Localization Investigations in a Parallel Intelligent Robot Cellular Automata Model that Performs Foraging Task", Proceedings of the 9th International Conference on Agents and Artificial Intelligence – Vol. 2: ICAART, pp. 355-363, 2017.

[4] J. Barraquand, and J. C. Latombe, "Robot motion planning: A distributed representation approach", Int. J. of Robotics Research, Vol. 10, pp. 628-649, 1991.

[5] O. Cliff, R. Fitch, S. Sukkarieh, D. Saunders, and R. Heinsohn, "Online Localization of Radio-Tagged Wildlife with an Autonomous Aerial Robot System", Proceedings of Robotics: Science and Systems, 2015.

[6] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-Robot Simultaneous Localization and Mapping: A Review", Journal of Field Robotics 33.1, pp. 3-46, 2016.

[7] T. Kuno, H. Sugiura, and N. Matoba, "A new automatic exposure system for digital still cameras", Consumer Electronics, IEEE Transactions on 44.1, pp.192-199, 1998.

[8] T. Sebastian, D. Fox, W. Burgard, and F. Delaert, "Robust Monte Carlo localization for mobile robots", Artificial Intelligence, Vol. 128, Iss. 1-2, pp. 99-141, 2001.

[9] W. Yunfeng, and G. S. Chirikjian, "A new potential field method for robot path planning", IEEE, DOI: 10.1109/ROBOT.2000.844727, San Francisco, CA, USA, 2000.

[10] D. Hahnel, W. Burgard, D. Fox, K. Fishkin, and M. Philipose, "Mapping and localization with RFID technology", *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, pp. 1015-1020, 2004.

[11] J. Flores, S. Srikant, B. Sareen, and A. Vagga, "Performance of RFID tags in near and far field", Proc. IEEE Int. Conf. Pers. Wireless Commun., pp. 353-357, 2005.

[12] D. Lima, and G. de Oliveira, "A cellular automata ant memory model of foraging in a swarm of robots.", Applied Mathematical Modelling 47, pp. 551-572, 2017.

[13] B. Siciliano, and O. Khatib, Handbook of Robotics, Ed. 2, ISBN: 978-3-319-32550-7, Springer –Verlag, Berlin, Heidelberg, 2016.

[14] J. Shi, and J. Malik, "Normalized cuts and image segmentation", Pattern Analysis and Machine Intelligence, IEEE Transactions on 22.8, pp. 888-905, 2000.

[15] R. Stengel, "Robot Arm Transformations, Path Planning, and Trajectories", Robotics and Intelligent Systems, MAE 345, Princeton University, 2015.

[16] S. Se, D. Lowe, and Jim Little, "Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks", The International Journal of Robotics Research, Vol 21, Iss. 8, pp. 735 – 758, 2002.

[17] T. Kuno, H. Sugiura, and N. Matoba, "A new automatic exposure system for digital still cameras", Consumer Electronics, IEEE Transactions on 44.1, pp.192-199, 1998.

[18] D. Hahnel, W. Burgard, and S. Thrun, "Learning compact 3D models of indoor and outdoor environments with a mobile robot", Robotics and Autonomous Systems, Vol. 44, Iss. 1, 2003.

[19] L. Lee, R Romano, and G Stein, "Introduction to the special section on video surveillance", IEEE Transactions on Pattern Analysis and Machine Intelligence 8, pp. 740-745, 2000.

[20] M. Beetz, "Plan-Based Control of Robotic Agents: Improving the Capabilities of Autonomous Robots", ISSN-0302-9743, Springer-Verlag, Berlin, Heidelberg, New York, 2002.

[21] C. Richter, S. Jentzsch, R. Hostettler, J. Garrido, E. Ros, A. Knoll, F. Rohrbein, P. van der Smagt, and J. Conradt, "Musculoskeletal robots: scalability in neural control", IEEE Robotics & Automation Magazine, Vol. 23, Iss. 4, pp. 128-137, 2016.

[22] D. Lima, and G. de Oliveira, "A cellular automata ant memory model of foraging in a swarm of robots.", Applied Mathematical Modelling 47, pp. 551-572, 2017.

[23] M. Alkilabi, A. Narayan, and E. Tuci, "Cooperative object transport with a swarm of e-puck robots: robustness and scalability of evolved collective strategies", Swarm Intelligence, Vol. 11, Iss. 3-4, pp. 185-209, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 394-397 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Elliptical Printed Dipole Antenna Design using ANN Based on Levenberg–Marquardt Algorithm

Ali I. Hammoodi[1*], Mariofanna Milanova[2], Haider Raad[3]

[1]*Department of Systems Engineering, University of Arkansas at Little Rock, 72204, USA*

[2]*Department of Computer Science, University of Arkansas at Little Rock, 72204, USA*

[3]*Physics Department, Xavier University, 45207, USA*

ARTICLE INFO

Article history:
Received: 11 August, 2018
Accepted: 08 October, 2018
Online: 18 October, 2018

Keywords :
Antenna Design
Flexible Antenna
Neural Network

ABSTRACT

A design of elliptical printed dipole antenna based on neural network approach is presented in this paper for recent WiMAX, Bluetooth, WLAN, LTE, and future 5G applications. The dipole patch is printed on top of substrate which has relative permittivity of 2.2 and 0.787mm of thickness. The elliptical dipole antenna is single band and the operational frequency is controlled by the major and minor axes of the ellipse. Neural network composed of one input layer, one hidden layer, and one output layer is implemented using simulated dataset of 24 samples collected using the electromagnetic simulator CST Microwave Studio. The neural network is trained using Levenberg-Marquardt algorithm and we divided the data set into 90% for training, 5% for testing, and 5% for validation. The error, described by the difference between the target data and expected output, is less than 0.005. The neural network model is a promising candidate to implement the design of elliptical dipole antenna to eliminate the simulation time taken by the EM simulator and reduce the cost of using large computing cluster.

1. Introduction

Single band antennas recently become more demanded designs for modern WLAN, WiMAX, and ISM applications due to the small size and uniform radiated gain with minimized side lobes. The light weight feature, low profile, flexibility, conformability, and ease of fabrication increased the demand of printed antennas in integrated circuit of modern wireless systems [1].

There are various tools of electromagnetic simulators have been used to design and analyze variety of antennas and built using computational electromagnetic analysis which aims to solve Maxwell equations for particular design. The three most common software's are: Ansoft HFSS, CST Microwave studio and FEKO. Some of these software's are based on Finite Element Method and some others are based on Finite Difference Time Domain and Finite Difference Frequency Domain as well as Method of Moment. The main disadvantages of these tools are, but not limited to: 1) Complexity of the structure determines the simulation time hence some of the structures could take up to several days , 2) The individual license cost of these tools are extremely high and could go up to thousands of dollars, and 3) requirements of powerful computing machines to handle the simulation structure of these tools and solve the design in considerable time. As a result of these major drawbacks, the approach we believe is the best candidate to overcome these issues is machine learning. This approach could be utilized to implement microwave circuit and antennas designs despite the fact of long simulation time and by this technique we could obtain cost-effective simulator.

There are different machine learning approaches could be used to implement EM simulators, however, Artificial Neural Network (ANN) is the highly recommend technique due to its capabilities in learning of solving complex and non-linear problems in different aspects. The weight of the neural network adjustment depends on the training algorithm and the given data set. Different training algorithms could be used which give different network performance and depends on the type of the problem we are trying to solve, the size of the training dataset and how we divide them to train the model and number of layers we could use to implement the model [2-4]. The technique of ANN has been implemented on several designs with different functionality. One example of E-shaped slot embedded in microstrip patch antenna where the neural network is used to obtain the effect of feeding point on the antenna performance [5]. Another example of C-Shaped patch antenna

*Corresponding Author: Ali Hammoodi, aihammoodi@ualr.edu

which is implemented using Bayesian Regularization algorithm to calculate the frequency of operation. A compact triple-band PIFA is implemented using ANN for antenna modeling [6].

This paper presents the design of elliptical shape printed dipole antenna using ANN. The ANN constructs of 1 hidden layers of 10 neurons. The Levenberg-Marquardt algorithm based on backpropagation is used to optimize the model. Parametric study and sensitivity analysis have been done on the design of the printed dipole to measure the effect of each design parameters on the frequency of operation. It is noted that the design of the printed dipole antenna is mostly affected by the major and minor axes of the elliptical shape and the gap between the two ellipses. Therefore, the input of the ANN is considered the desire frequency of operation and the output represented by the major axis of the ellipse, minor axis of the ellipse, and the maximum expected gain at the design frequency. The ANN is train using dataset of 24 samples. The dataset was collected using CST Microwave studio, 2017 [7]. The optimized ANN model is achieved when the dataset is split as 90% for training set, and 5% for both validation and testing. The error is less than 0.005 in all the phases as this error representing the difference between the output of the ANN and the expected value. The ANN presented in this paper is a good approach for the design of single band dipole antenna for modern wireless applications.

The follow sections are divided and summarized as: Section two possess antenna geometry and the antenna performance under one case study. Section three describes the neural network model and analysis. Section four concluded the presented work and followed by the used references.

2. Antenna Geometry and Design

The printed dipole antenna is modeled using CST microwave studio, 2017. The scattering parameters (S-parameters) is used to show the performance of the model using EM simulator. Dipole antenna is built on substrate of 2.2 relative permittivity and 0.787 mm of thickness. The dipole is fed at the center by discrete face port. Figure 1 shows the geometry of the printed elliptical dipole.

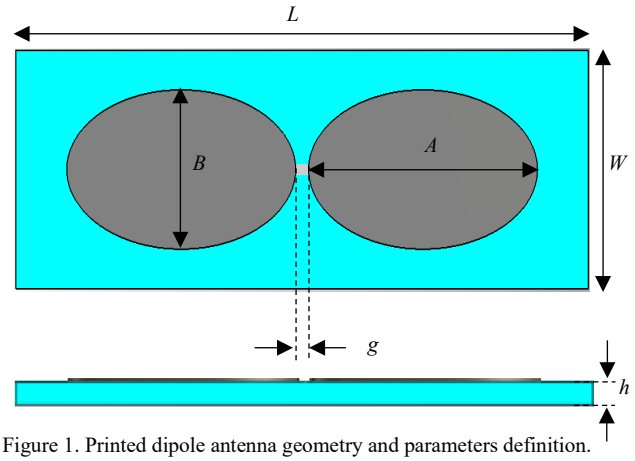


Figure 1. Printed dipole antenna geometry and parameters definition.

The numerical method implemented in the CST EM tool is based on Finit Difference Time Domain which is used to solve Maxwell equations for antennas using time domain analysis. The operational frequency of antenna is represented by the S_{11} and which represemts the ratio of the reflected signal to the incident signal. By understanding the design philosophy using sensitivity analysis and parameteric study, the operational frequency of the given printed elliptical dipole antenna is controlled by the major and minor axes of the ellipctcal shapes and the gap between the two ellipes is used to achieve impedance matching. One case study of S-parameters is shown in Figure 2 and the design parameters are given in Table 1.

The operational frequency of antenna is given by the point where the S_{11} is ≤ -10 dB. This criterion is a must to define valid data set hence the leak of the previous research is addressed by using data set with invalid data points to train the network. As shown in Figure 2, the resonance frequency (f_o) is equal to 6.28 GHz.

Table 1. Design parameters of the case study

Parameter	Value	Parameter	Value
L	27.5 mm	A	11 mm
W	8.25 mm	B	5.5 mm
g	0.624 mm	h	0.787 mm

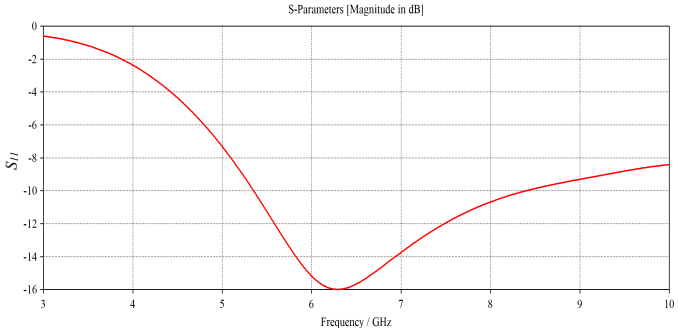


Figure 2. S-parameters of the elliptical shape printed dipole.

3. Neural Network Model

There are several undesired factors encouraged the use of neural network approach for the design of antennas and 3D modeling including, the time the simulator takes to solve the structure, the need of powerful computing unit, and cost of purchasing software's license. As mentioned eariler, machine learning, specifically ANN, is considered as a good candidate to synthesis the design parameters and performance metric of antennas without the need of expesive EM tools. Moreover, implementing designs using 3-D field simulators requiring the understanding of the EM boundaries and requirements for the basic by which the software is built to maintain realistic simulation enviroment.

In this research paper, neural network is used to model the design synthesis of the elliptical printed dipole antenna. The ANN model presented in this paper is built using dataset composed of the S_{11}, realized gain, and antenna's geometry. The implemented network is build of single input layer of one input parameter which is the antenna operating frequency, one optimzied hidden layer of 10 neurons and the output layer is composed of four variables gain (G), major axis (A), minor axis (B), and the gap between the two ellipse (g) as shown in Figure (3).

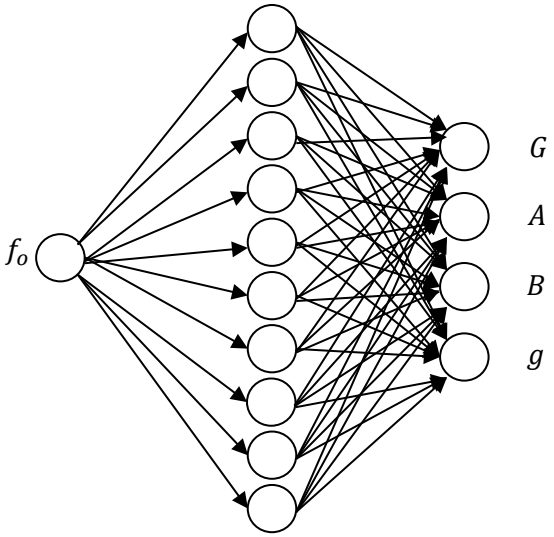


Figure 3. The model of the ANN for the elliptical dipole antenna.

The desired operational frequency is the input to the nerual network and the goal of the ANN is to synthesis the the physical dimenions and the realized gain. The simulated data of 24 samples were collected using the well-known EM simulator CST Microwave Studio, 2017. The major factor to make the ANN model succeed in antenna design is to have valid collected data where the requirement of S-parameters ≤ -10 dB is considered. The goal of the implemented ANN is to predict the physical dimenions of the dipole antenna working within the (1-10) GHz as most of the current and modern application work in this frequency band.

The ANN network used is based on the backpropogation technique and the Levenberg-Marquardt algorithm has been used to train the ANN and adjust the weights to predict the parameters of the elliptical printed dipole antenna based on the given operational frequency. In the hidden layer, the neurons compute its trigger value using the activation fucntion given in (1) which shows that each value of basis function for each neuron is a result of multiplying the activation function by the sum of each input multiplies by the weight [2]:

$$\varphi_i(x) = \sigma \sum_k a_{ij} x_k + b_i \quad (1)$$

where, φ_i is the basis function, σ is the activation function (which is a sigmoid function used in this paper), a_{ik} is the node weight, and x_k is the input value. The system output (y) is described by (2):

$$y = \sum w_i \varphi_i \quad (2)$$

where, w_i is the orthogonal weight. MATLAB is software utilized to model the network. The dataset is divided into 90% for learning, and 5% for both validation and testing. The output regression of the ANN is shown in Figure 4 and the resultant prediction rate is more than 99%. The regression is described by how the predicted output of the neural network in difference phases are align with the given training data after optimizing the weights. The good regression is when the data are aligned along the slop of the regression (Fit).

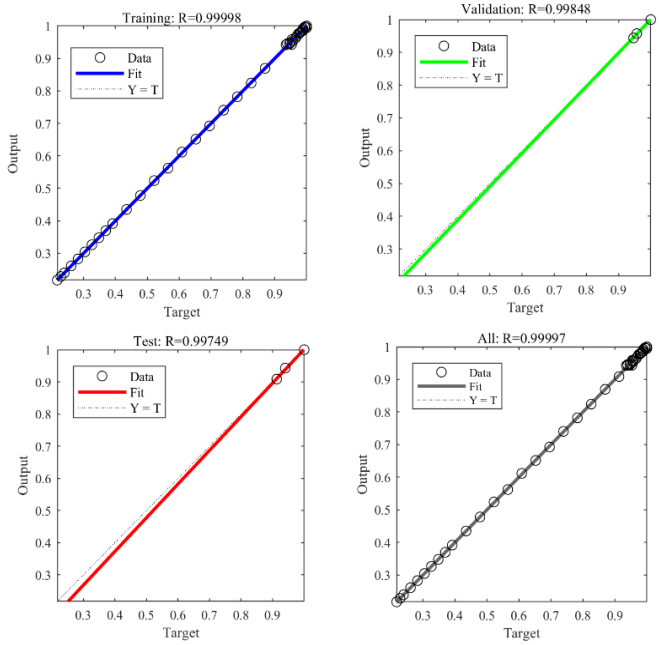


Figure 4. The regression of the neural network in training, validation, and testing phases.

It is noted that the amount of error given by the difference between the output of the ANN and the dataset shown in Figure 5 is below 0.005 over the entire dataset. Therefore, the data set of 24 samples is sufficent to learn the ANN hence the neural network discribed in this paper is a promising candidate for modeling the elliptical printed dipole antenna.

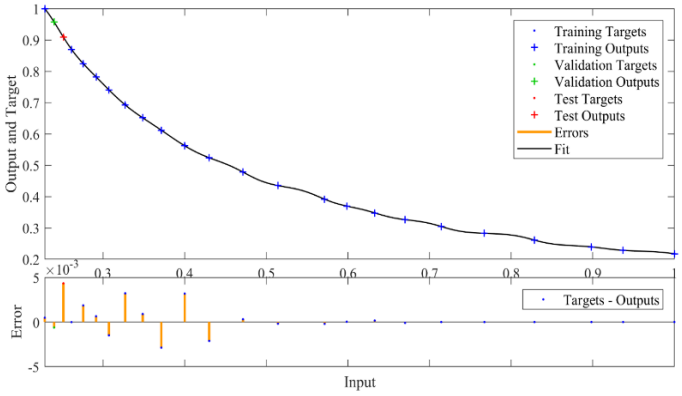


Figure 5. The error between the predicted output and the target output given in the data set.

4. Conclusion

This paper presents the design of elliptical printed dipole antenna using machine leanring based neural network. The dipole is printed on top of substrate with 2.2 of relative permittivity and 0.787 mm of thickness. The dipole is fed at the center point by discrete port. Based on sensitivity analysis, the three design parameters control the operation frequency of the dipole; major and minor axis and the gap between the two elliptical poles. The dataset is collected using the CST microwave studio, 2017. The the neural network singe input is the desire frequency of operation and the target output is defined by the physical dimentions and the predicted gain. The traing algorithum is Levenberg-Marquardt

which is based on backpropagation. The data set is divided into 90% for learning, and 5% for both validation, and testing. The error difference between the output of the trained network and the target output is less than 0.005. The ANN is highly recommended candidate for the design of elliptical printed dipole single band antenna to avoid the time consumed using EM tools, and the requirement of powerful computing servers.

References

[1] C. A. Balanis, *Antenna Theory: Analysis and Design*. Wiley, 2016.
[2] Livingstone, D.J. Artificial Neural Networks: Methods and Applications. Humana Press, New York, 2009.
[3] V. S. Kushwah and G. S. Tomar, "Design and Analysis of Microstrip Patch Antennas Using Artifical Neural Network," in Trends in Research on Microstrip Antennas, InTech, 2009, pp. 720–724.
[4] V. V Thakare and P. Singhal, "Microstrip Antenna Design Using Artificial Neural Networks," Int. J. RF Microw. Comput. Eng., pp. 210–224, 2009.
[5] G. Singh and S. Kaur, "Design Analysis of an E-Shaped Slot Loaded MPA and Parameters Estimation using ANN," in *International Conference on Computing, Communication, and Automation (ICCCA2016)*, 2016, pp. 1404–1408.
[6] R. Vanna and J. Ghosh, "Modeling of a compact triple band prF A using Knowledge based neural network," in *2016 IEEE Region 10 Conference (TENCON)*, 2016, pp. 2190–2192.
[7] CST Microwave Studio, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 398-406 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

The Effect of Fluence Variations of Nd:YAG Laser Ablation and Sample Condition on Human Tooth

Fatanah Mohamad Suhaimi*[1], Nur Zarifha Zainol Alam[2], Suriani Mat Ariffin[2], Nurul Atiqah Abd. Razak[1], Mohammad Khairul Azhar Abdul Razab[3]

[1]*Craniofacial and Biomaterial Sciences Cluster, Advanced Medical and Dental Institute, Universiti Sains Malaysia, Bertam, 13200 Kepala Batas, Penang, Malaysia*

[2]*School of Physics, Universiti Sains Malaysia, 11800 Gelugor, Penang, Malaysia*

[3]*School of Health Science, Universiti Sains Malaysia, Health Campus, 16150 Kubang Kerian, Kelantan, Malaysia*

ARTICLE INFO

Article history:
Received: 14 August, 2018
Accepted: 11 October, 2018
Online: 18 October, 2018

Keywords:
Enamel
Laser ablation
Nd:YAG laser

ABSTRACT

Nd:YAG laser has shown some potential to be used in dental practice replacing the conventional method. In particular, it can be used to modify the tooth surface by the ablation process. The laser provides an ability to accurately deliver a significant amount of energy into a confined region. Thus, alteration of the sample surface and composition may occur during the process. Additionally, the use of a laser in ablation procedure is also associated with heat generation and potential thermal injury that may be experienced by a patient. The wet condition of the samples is expected to reduce the thermal effect. In this paper, the changes of enamel surface and elemental composition following laser irradiation of Nd:YAG laser are discussed. The teeth samples were irradiated at 1.5 Hz of pulse rate, 100 ms of pulse width, and range of fluences of 80-120 J/cm^2. Field Emission Scanning Microscopy (FESEM) and Energy Dispersive X-ray (EDX) were used to analyze the morphology and composition of the teeth samples. Samples were compared before and after laser irradiation. The percentages of carbon increased after laser irradiation, while oxygen decreased for most of the samples. The morphologies of the samples were varied with a more pronounced effect on the sample surface at higher fluence. In addition, the effect of wet sample condition is also investigated and discussed. It is demonstrated that the ablation in wet condition produced less damage to the enamel surface compared to dry sample. However, no remarkable difference between the elemental composition of wet and dry samples.

1. Introduction

Laser has been successfully used in dental applications such as for caries prevention, etching bond strength, restorative procedure, gum recontouring, and surgical procedure [1-3]. Research on medical laser in dental practice has been conducted for several years to investigate the potential and suitability of laser devices and its parameters for replacing the conventional method. However, careful consideration has to be made on selecting parameter settings, types of laser, wavelength, energy, pulse duration, absorption properties and scattering effect. In particular, the interactions of the laser with tissues are pivotal for selecting laser system and operating procedure that will best suit the application in dentistry.

A laser is categorized into four types depending on the medium used, whether it is a liquid, solid-state, gas and semi-conductor. Among all types of laser, Nd:YAG is commonly used in dentistry. The first pulsed Nd:YAG laser was released in 1990 specifically for the dental market [1]. Common laser parameters include fluence, pulse width and pulse rate. Fluence is the optical energy per unit area. Whereas, the pulse width can be defined as the amount of time the laser releases its energy or emits the laser pulse. Pulse rate indicates the number of pulses emitted per second.

In dentistry, Nd:YAG laser has been used for caries prevention [3], germ removal, and modify the bond strength and surface treatment [4, 5]. Nd:YAG laser possesses a suitable wavelength for

*Fatanah M. Suhaimi, Craniofacial and Biomaterial Sciences Cluster, Advanced Medical and Dental Institute, Universiti Sains Malaysia, 13200 Kepala Batas, Penang, Malaysia. +6045622561, fatanah.suhaimi@usm.my

high diffusion through hydroxyapatite and water, and soft tissue due to its behaviour of having high penetration depth [6]. Additionally, pre-treatment by Nd:YAG laser irradiation of etch-and-rinse adhesive appeared to have a positive effect on the adhesive-dentin bonding [5]. An in-vitro study by Beketova et al. suggested that Nd:YAG laser has potential in inducing bioactivity of dental ceramic [7].

Eventhough many research had shown the positive effect of using Nd:YAG laser, the impact toward the exposed tissues need to be considered to avoid damage and deterioration of the tissue. Issues related to the ablation process of dental tissue include irregular surface modification, charring, and peripheral cracks [8]. These issues are also associated with dry ablation. An alternative method by lasers for dental ablation characterize the risk involved such as thermal damage experience by patient compare to traditional cavity preparation methods. Moreover, heat generation is crucial as it may lead to painful sensations and potential thermal injury [8-11].

Laser ablation requires sufficient energy to remove material from the tooth by ejecting the electrons from its atom. A study by Cecchini et al. reported that dimineralization of enamel occurs after laser irradiation. Change in morphology is indicated with cracks and crater presence after irradiation [12]. In addition, a research conducted by Khatavkar and Hegde found that Er:YAG laser cause an etching effect on a surface that is comparable to conventional acid etching [13]. Increasing the energy parameters produced a difference surface morphology of enamel from roughening to an etching-like micro-roughen pattern based on a Field Emission Scanning Electron Microscope (FESEM) evaluation.

Histological evaluation on the efficiency of employing ultrashort pulsed laser in dentin and enamel caries removal provided the positive result as no morphological changes and no side effect indicated in the healthy areas of dentin and enamel [8-9]. On the other hand, morphological studies on enamel and dentin irradiated with Nd:YAG laser with different energy densities shows that enamel exhibits shallow crater than dentin due to the fact of varying composition between them. Enamel tissues are found to be more calcified, tight and hard which promotes fissures due to sudden heat distribution throughout the surface. However, the high number of shots resulted in significant changes in morphology with total eradication of smear layer [14].

Besides, the elemental composition of the enamel may also be altered due to the laser irradiation subjected to the sample. According to Al-Hadeethi et al., 2016, radiation by high energy laser melts the enamel surface, which then increased energy density, and thus increases carbon element [15]. Moreover, photothermal evaporation during laser irradiation caused a reduction of water molecules [16]. The changes in water molecules may be assessed by examining the oxygen composition. Interestingly, calcium, phosphorus, and the Ca/P ratio in the sample indicate the hardness of the tooth.

The condition of the sample whether dry or wet plays an important role towards the surface produced after the ablation process. A study by Bartoli et al. on stone cleaning application indicates that ablation in wet condition is more efficient due to vaporization process of water molecules [17]. This statement is also supported by a study conducted on a wet film that suggests the wet condition has less resistant to the laser irradiation [18]. Thus, thermal conductivity, heat transfer and vaporization process that occur during the ablation process may result in different morphological changes of the tooth samples.

This research was done to investigate the effect of Nd:YAG laser fluences on the enamel of the human tooth. Other laser parameters such as pulse width and pulse rate were kept constant. Surface morphology and elemental composition of the enamel were analysed before and after laser irradiation. Additionally, the condition of the sample whether wet or dry is also examined to identify the differences outcome of the laser ablation process. This paper is an extension of work originally presented in the 39th International Conference of IEEE Engineering in Medicine and Biology Society [19].

2. Methods

2.1. Sample Preparation

In this study, seven premolar human teeth were used as the sample. Teeth were collected from the dental clinic of the Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia (USM) in Penang, Malaysia. The collected samples were stored individually in a plastic container containing normal saline. Samples were carefully selected, and only healthy teeth were used for the study. Ethical approval for this study has been obtained from the Human Research Ethics Committee of Universiti Sains Malaysia.

Debris on the teeth was removed using a brush and cleaned using water. Then, the cleaned tooth was cut into two using an IsoMet Low-Speed Saw, (Buehler, US). After that, the tooth was mounted in a designated container, and the surface of the mounting was covered with a mixture of resin and hardener. The amount of the resin and the hardener were measured carefully by volume. The sample surface was then polished using a grinding and polishing machine, MetaServ 250 Grinder Polisher (Buehler, Germany). During the grinding procedure, a rotating disk of abrasive paper was used, and it was flushed simultaneously with a coolant to remove debris and minimize the heat.

2.2. Laser Irradiation

Seven premolar teeth were used in this study. Samples were irradiated with Nd:YAG laser (Cynosure, US) at a wavelength of 1064 nm. A 5 mm laser tip was used. The teeth were irradiated with a fluence of 80, 90, 100, 110, and 120 J/cm^2, a pulse rate of 1.5 Hz, and a pulse width of 100 ms. One sample was irradiated for each fluence. Thus, a total of five samples were irradiated in dry condition. The distance of the laser tip was kept constant during the irradiation procedure to ensure a consistent spot size on the enamel surface.

2.3. Sample Analysis

Electron Dispersive X-ray Spectrometer (EDX) and FESEM were used to analyse the samples. The image of the sample was captured, and the morphological analysis was made using the FESEM at several magnifications. The elemental composition in terms of atomic percentage and weight percentage was recorded using the EDX system. The morphology and data on the elemental

composition of the samples were compared before and after laser irradiation.

2.4. Sample Condition

Wet samples were also tested to investigate whether the samples produce different morphology compared to the dry samples. Water spray was applied onto the sample surface before the irradiation process. Two samples were irradiated in the wet condition.

3. Results and Discussion

3.1. Effect of Fluence Variations

The morphology of the enamel surface had been captured before and after laser irradiation with different fluence energies applied to each sample. Figure 1 shows the surface morphology of tooth sample before and after laser irradiation with 80 J/cm^2 at 5000x magnification. The surface appeared rough and non-uniform with large bubbles present after irradiation as compared to before irradiation.

Table 1 presents the identified elements by atomic and weight percentages before and after laser irradiation at 80 J/cm^2. In this table, C is carbon, O is oxygen, Na is sodium, P is phosphorus, Cl is chlorine, Ca is calcium, and Si is silicon. Carbon had the highest atomic percentage after irradiation (54.38%) followed by oxygen (28.20%). Similarly, the weight percentages were 37.78% and 26.10%, respectively. Interestingly, oxygen was the highest element before irradiation. Elements such as Na, P, and Cl were not largely affected after laser irradiation since the atomic and weight percentages were muchly smaller compared to other elements.

Table 1: Identified elements by atomic percentage and weight percentage, before and after laser irradiation at 80 J/cm^2.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	13.31	7.24	54.38	37.78
O	53.87	39.02	28.20	26.10
Na	0.65	0.68	0.40	0.54
P	12.48	17.51	6.09	10.92
Cl	0.70	1.13	0.62	1.25
Ca	18.97	34.42	9.61	22.27
Si	-	-	0.70	1.14
Total	100.00	100.00	100.00	100.00

Figure 2 shows the surface morphology of the tooth sample before and after laser irradiation with 90 J/cm^2 at 5000x magnification. The image consists of large indentations with less bubbles evenly disperse throughout the surface. However, more bubbles were present, and the surface appeared rough after being irradiated by laser at 90 J/cm^2, indicating that the resulted surface had more damages as compared to the sample in Figure 1. However, the surface of the sample before irradiation as shown in the left panel appeared scalier compared to the first sample in Figure 1.

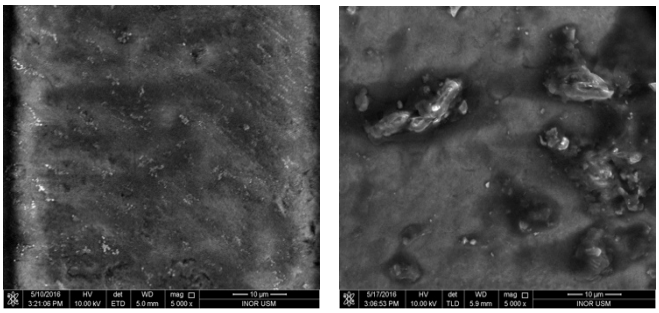


Figure 1: Surface morphology of tooth sample before (left) and after laser irradiation (right) of 80 J/cm^2 at 5000x magnification.

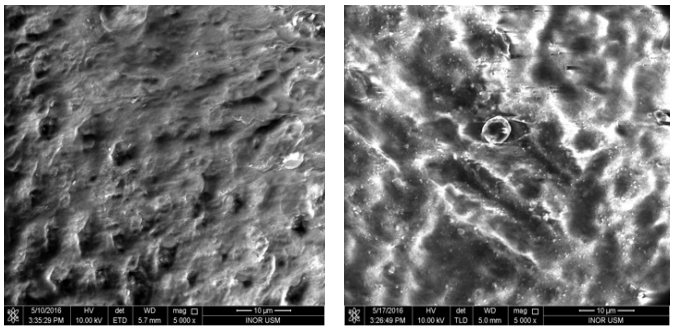


Figure 2: Surface morphology of tooth sample before (left) and after laser irradiation (right) of 90 J/cm^2 at 5000x magnification.

Table 2 presents the identified elements by atomic and weight percentages before and after laser irradiation at 90 J/cm^2. Oxygen, phosphorus, and calcium decreased in both atomic and weight percentages. The percentage difference of weight changes were less than 12% individually for oxygen, phosphorus, and calcium. The highest element after irradiation was oxygen, eventhough oxygen slightly decreased after irradiation. Nevertheless, carbon increased after the irradiation, from 21.33% to 38.85% of atomic percentage. Chlorine and silicon were not extremely modified by the irradiation.

The surface morphology of tooth sample before and after laser irradiation with 100 J/cm^2 at 5000x magnification is shown in Figure 3. Enamel surface appears smooth and shiny with evenly choppy texture. Nonetheless, the appearance of the enamel changed after laser irradiation resulting in a rough, and an uneven bubble distribution on the surface of the sample.

Table 3 presents the identified elements by atomic and weight percentages before and after laser irradiation at 100 J/cm^2. In this table, Al is aluminium. Oxygen, phosphorus and calcium decreased significantly while carbon increased by almost triple the value before irradiation. The atomic percentage of oxygen was largely reduced, which was from 52.26% to 36.68%. Other elements were not significantly changed.

The tooth sample before and after laser irradiation with 110 J/cm^2 at 5000x magnification is shown in Figure 4. The surface appears uniform and rough with fewer indentations before laser irradiation. However, cracks and non-uniform surface were presented after exposing the sample to laser irradiation.

Table 2: Identified elements by atomic percentage and weight percentage, before and after laser irradiation at 90 J/cm^2.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	21.33	12.44	38.85	25.13
O	51.89	40.31	40.01	34.48
Na	-	-	0.74	0.92
P	10.21	15.36	7.57	12.63
Cl	0.42	0.71	0.53	1.01
Ca	15.70	30.56	11.20	24.17
Si	0.45	0.62	1.10	1.66
Total	100.00	100.00	100.00	100.00

Table 3: Identified elements by atomic percentage and weight percentage, before and after laser irradiation at 100 J/cm^2.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	18.66	10.63	44.37	30.19
O	52.26	39.63	36.68	33.25
Na	0.63	0.68	0.55	0.71
P	11.57	16.99	7.66	13.45
Cl	-	-	0.48	0.96
Ca	16.88	32.07	7.59	17.24
Si	-	-	1.91	3.04
Al	-	-	0.76	1.16
Total	100.00	100.00	100.00	100.00

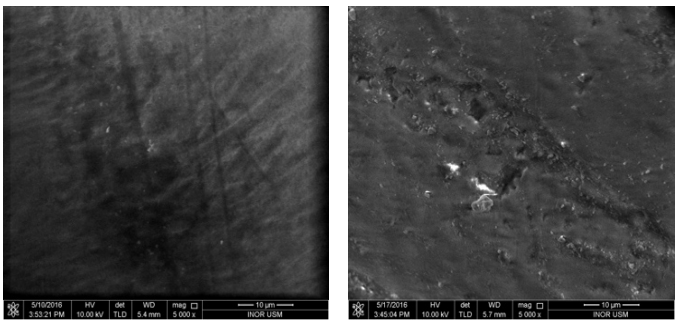


Figure 3: Surface morphology of tooth sample before (left) and after laser irradiation (right) of 100 J/cm^2 at 5000x magnification.

Table 4 presents the identified elements by atomic and weight percentages of the tooth sample before and after laser irradiation at 110 J/cm^2. The percentages of carbon, oxygen, and sodium reduced after laser irradiation. However, the change in oxygen was very light compared to the changes in carbon. Both atomic and weight percentages of phosphorus and calcium were increased. Yet, the changes in calcium were profound compared to phosphorus. Chlorine and silicon were not detected before irradiation but appeared after irradiation with a very small percentage.

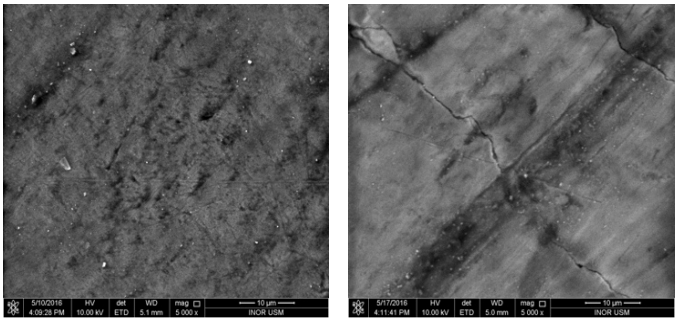


Figure 4: Surface morphology of tooth sample before (left) and after laser irradiation (right) of 110 J/cm^2 at 5000x magnification.

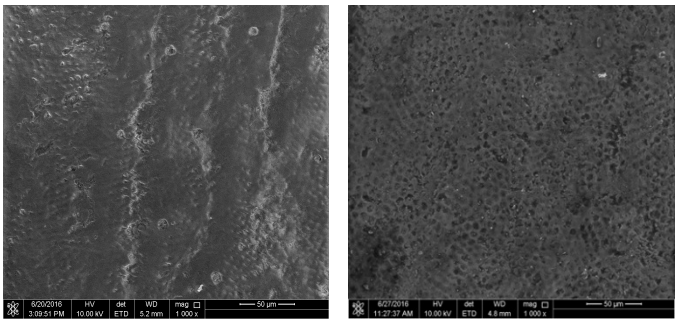


Figure 5: Surface morphology of tooth sample before (left) and after laser irradiation (right) of 120 J/cm^2 at 1000x magnification.

Figure 5 shows the surface morphology of a tooth sample before after laser irradiation at 120 J/cm^2 at 1000x magnification and 500x magnification, respectively. The surface of the sample before irradiation was scaly and uneven. Nevertheless, the surface appeared even with tiny hollows after laser irradiation.

Table 4: Identified elements by atomic percentage and weight percentage, before and after laser irradiation at 110 J/cm^2.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	35.76	23.68	26.18	15.21
O	45.66	40.28	45.26	35.01
Na	0.55	0.70	0.41	0.45
P	8.94	15.26	11.19	16.76
Cl	-	-	0.32	0.56
Ca	9.09	20.08	16.23	31.45
Si	-	-	0.41	0.56
Total	100.00	100.00	100.00	100.00

Table 5 compares the elemental compositions of a tooth sample before laser irradiation and after laser irradiation of 120 J/cm^2. In this table, Mg is magnesium. Before irradiation, oxygen has the highest atomic percentage with 38.29%, followed by calcium with 28.15%, phosphorus with 16.38%, and carbon with 16.24%. The

highest weight percentage is calcium with 45.68%, followed by oxygen, phosphorus, and carbon. Carbon and oxygen increased after laser irradiation. However, other elements decreased after laser irradiation. The percentages of silicon, sodium, magnesium and chlorine were considered not significant compared to other elements in this sample.

Table 5: Identified elements by atomic percentage and weight percentage, before and after laser irradiation at 120 J/cm^2.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	16.24	7.90	35.70	22.32
O	38.29	24.81	40.74	33.92
Na	0.54	0.50	0.30	0.36
P	16.38	20.54	8.52	13.72
Cl	0.40	0.57	0.37	0.69
Ca	28.15	45.68	12.84	26.79
Si	-	-	1.33	1.95
Mg	-	-	0.20	0.25
Total	100.00	100.00	100.00	100.00

3.2. Effect of Wet Sample Condition

Figure 6 shows the surface morphology of the tooth sample before and after laser irradiation of 80 J/cm^2 at a wet condition. The surface after laser irradiation at a wet condition appeared more uniform with tiny holes presented. Additionally, Table 6 compares the identified elements by atomic percentage and the weight percentage of the tooth sample before and after laser irradiation of 80 J/cm^2 at the wet condition. In this table, C is carbon, O is oxygen, Na is sodium, P is phosphorus, Ca is calcium, Si is silicon, and Mg is magnesium.

Oxygen was the highest element before irradiation, followed by calcium and phosphorus. After laser irradiation, the atomic percentage of oxygen reduced significantly while slight changes can be observed in the weight percentage. On the other hand, carbon increased massively to almost six times of its atomic percentage and nine times of its weight percentage, becoming the highest element exist in the sample. Phosphorus and calcium were also decreased after laser irradiation. Silicon was not detected before irradiation, but appeared after irradiation with a very small percentage. Nevertheless, sodium and magnesium were not found in the sample after irradiation.

Figure 7 presents the surface morphology of a tooth sample before and after laser irradiation of 110 J/cm^2 at a wet condition with a magnification of 1000x, and 500x, respectively. The sample appears uniform, smooth and without crack formation. Additionally, Table 7 shows the identified elements by atomic percentage and weight percentage before irradiation and after laser irradiation of 110 J/cm^2 at the wet condition.

Oxygen was the highest element in the sample before irradiation and after laser irradiation, eventhough it was largely decreased after irradiation. Additionally, carbon also decreased after laser irradiation. Other elements include sodium, phosphorus, and calcium were increased after laser irradiation. However, the increased in calcium was significant compared to the changes in phosphorus and sodium. Chlorine and silicon were not detected before irradiation but appeared after irradiation with a very small percentage.

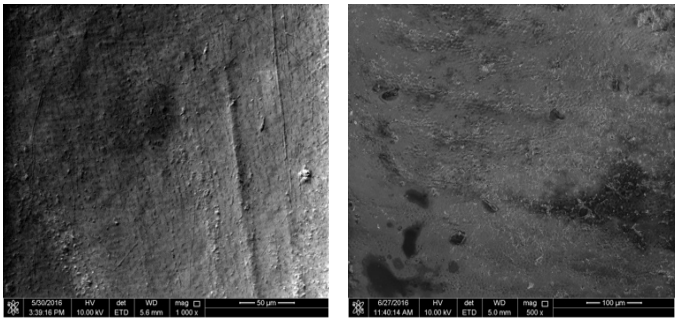


Figure 6: Surface morphology of the tooth sample before laser irradiation (left) at 1000x magnification and after laser irradiation with 80 J/cm^2 at wet condition (right) at 500x magnification.

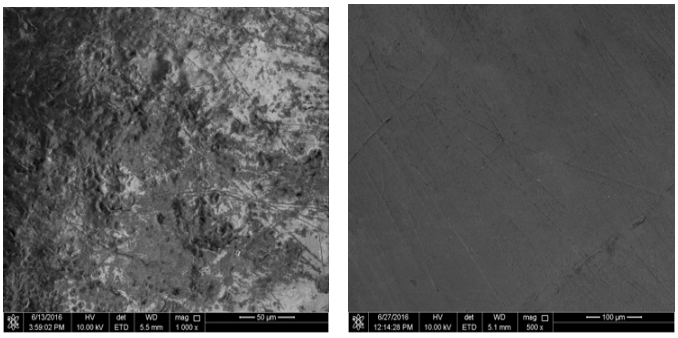


Figure 7: Surface morphology of the tooth sample before laser irradiation (left) at 1000x magnification and after irradiation with 110 J/cm^2 at wet condition (right) at 500x magnification.

Table 6: Identified elements by atomic percentage and weight percentage, before and after laser irradiation of 80 J/cm^2 with wet condition.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	8.66	4.87	49.12	38.46
O	62.60	46.88	43.59	45.46
Na	0.54	0.59	-	-
P	11.90	17.26	4.28	8.63
Ca	16.06	30.14	2.47	6.47
Si	-	-	0.54	0.98
Mg	0.24	0.26	-	
Total	100.00	100.00	100.00	100.00

Figure 8 presents the bar charts on atomic and weight percentage of C, O, P and Ca before irradiation on the top panel and after laser irradiation of 80 J/cm^2, 90 J/cm^2, 100 J/cm^2, 110 J/cm^2, and 120 J/cm^2 on the bottom panel. These four elements

were selected for further comparison across samples as they have more significant values, highly associated with the ablation process, and important element for tooth sample. In this figure, C is carbon, O is oxygen, P is phosphorus, and Ca is calcium.

Table 7: Identified elements by atomic percentage and weight percentage, before and after laser irradiation of 110 J/cm^2 with wet condition.

Element	Before Irradiation		After Irradiation	
	Atomic	Weight	Atomic	Weight
C	22.59	14.93	18.57	10.18
O	60.93	53.63	48.44	35.35
Na	0.50	0.63	0.51	0.54
P	8.84	15.07	12.31	17.40
Cl	-	-	0.55	0.89
Ca	7.14	15.74	19.20	35.11
Si	-	-	0.42	0.53
Total	100.00	100.00	100.0	100.00

There is an abundance of oxygen concerning the atomic and weight percentages before irradiation as observed in Figure 8, followed by carbon, calcium, and phosphorus. However, the percentages of carbon and calcium were varied among these samples, whereas the percentages of phosphorus are about the same. The weight percentage of carbon following irradiation increased tremendously, except for the sample irradiated with a fluence of 110 J/cm^2. After laser irradiation, the percentages of carbon were significantly increased for most of the samples while calcium and oxygen reduced profoundly. Additionally, phosphorus is not majorly changed by the laser irradiation.

In general, the increased and decreased pattern of the four elements with regards to the atomic percentage is about the same across samples. Yet, the percentage differences are highly variable from one sample to another. Only tooth sample that has been irradiated with 110 J/cm^2 has an opposite pattern. These patterns of element changing are almost similar for all the samples presented regardless of the fluence used during the ablation process.

Figure 9 presents the bar charts on atomic and weight percentages of C, O, P and Ca before irradiation and after irradiation with 80 J/cm^2 and 110 J/cm^2 at wet condition. The trend of tooth sample irradiated with 80 J/cm^2 is similar to the overall trend seen in Figure 8. Nevertheless, the trend of tooth sample irradiated with 110 J/cm^2 is contrary.

Comparing the same fluence used at 80 J/cm^2 but at different sample conditions, wet and dry, the percentage of carbon increased almost similar. However, the oxygen content of the dry sample is relatively higher compared to the dry sample. However, the changes in oxygen before and after irradiation for both conditions are about the same. In contrast, the percentage difference of calcium before and after irradiation was less for the wet sample. Thus, calcium is not muchly modified by the laser irradiation in the wet condition.

The samples irradiated with 110 J/cm^2 resulted in a lower percentage of carbon either at the wet or dry condition. The percentage of oxygen also decreased eventhough the wet sample has a larger percentage difference before and after laser irradiation. The percentages of phosphorus and calcium are not much different with regards to the wet or dry condition of the sample.

Theoretically, the higher the energy applied to the samples, more interactions occur giving more energy to the atom, and thus resulted into modifications in both morphology and mineral composition. Formation of cracks, shallow pit and fissures can be seen after higher fluence was applied. Nevertheless, wet samples appear uniform, flat, smooth and less crack formation as shown in Figures 6-7. Shallow indentation appearance on dry sample provides information on the rough surface caused by the laser with huge bubbles presented.

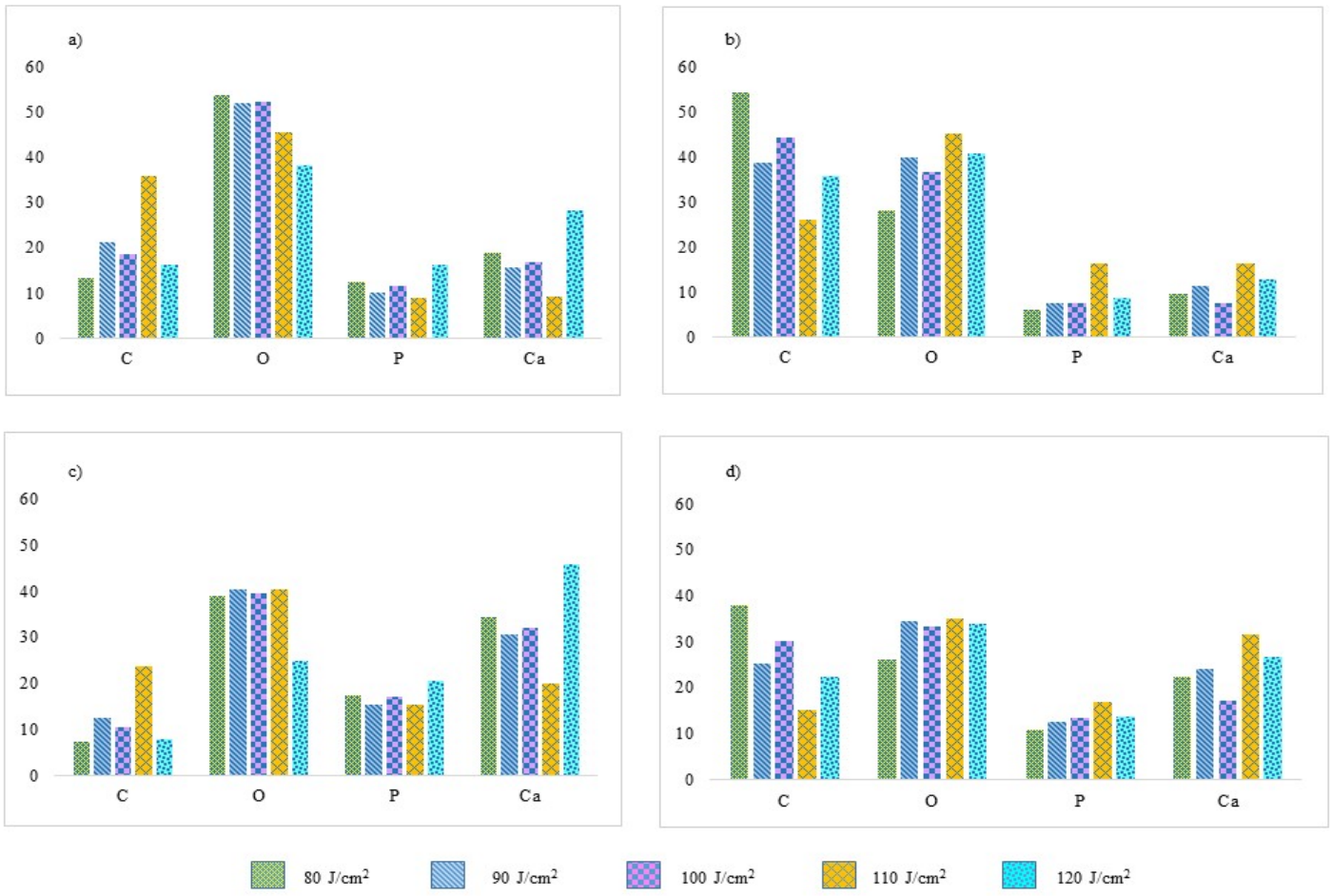
Inconsistent results of the tooth elements modifications show that the effect of the laser varies. Technically, it relies on the laser parameters include fluence energy, pulse width and pulse rate. However, the teeth samples used also affect the outcome. Despite the difference fluence used during ablation, the surface morphology and elemental composition of the collected teeth samples were also varied. Thus, the outcome is also expected to be varied, and the difference in the percentages of the elemental composition are assessed to conclude.

Carbon undergoes percentage increment after laser interaction with the samples in almost all fluence energies applied except irradiation at 110 J/cm^2. This explains the undesired carbonization phenomena within the tissues. As the temperature increases, the tissues tend to experience dehydration and eventually burned. Based on the ablation methodology, the absorption of laser beam energy process occurs on the tooth surface makes the surfaces become more carbonize [15, 20].

In contrast, carbon composition of the sample irradiated with 110 J/cm^2 experienced reduction, which is unexpected. This might due to the natural condition of the tooth sample. The carbon content of a tooth sample is correlated to several factors include enamel maturity, diets intake and caries susceptibility [21]. Increase caries susceptibility has been associated with the increase in carbon element [22]. Additionally, hypoplastic enamel contains more carbon compared to well-calcified enamel. Thus, the collected samples may vary in terms of enamel density, the number of enamel layers and the weight of enamel layers. In particular, the variation in enamel density results into the variation of optical property and laser absorption of the enamel layer. Thus, the contradict changes in carbon percentage is multifactorial and possibly due to the optical property of the sample itself.

In the case of the sample that has been exposed to a fluence of 80 J/cm^2, the carbon content appears to be the highest, even after the lowest fluence exposure. Photothermal interactions of the laser with tissues include absorption, penetration, scattering and reflection. The undesired carbonization occurs due to the higher interactions with biochemical compounds in the tissue sample. Similarly, this is potentially due to the enamel density and optical property of the sample.

Although the elemental composition might not largely affect, the surface morphology seems to have better form when the



surface is in the wet condition during laser irradiation. The wet condition of the samples changes the morphological appearance of a tooth surface as shown in Figures 6-7. Rough texture resulted from the dry sample leads to heat generation and thus charring pattern on the tooth surface. At a higher laser fluence, tooth surface may appear with a burning effect and peripheral cracks.

The application of water spray has been associated with low laser-ablation efficiency, as reported in several studies. The use of water spray also ensures the rehydration of mineral contains in the tissue [23]. This is particularly important for Nd:YAG laser ablation in which the influence of the water layer on maintaining the mineral content while preventing stalling and excessive peripheral damage. However, for dental application, the use of water spray is necessary to avoid heat sensation to the patients. A fluence up to 100 J/cm^2 seems acceptable to be used in addition to the wet condition of the sample, as seen from this study.

There are some limitations in this study. It is almost impossible to gather samples that have the similar amount of elemental composition. The compositions are highly variable among all the samples, and some other confounding factors such as enamel density, maturity and calcification may contribute toward the variation. Moreover, morphological appearance of one sample to another was also varied. Some of the samples may have been through a dental treatment which changed the morphological and elemental composition. Thus, comparison of the morphology and elemental composition has to be assessed based on the changes before and after laser irradiation.

As a suggestion, a larger number of samples can provide more statistically significant of the reduction and increment of the elemental composition. Selecting teeth sample from the same subject may reduce the variation. Additionally, the thickness of the sample can be standardized to minimize enamel layer variation, and topography analysis have to be carried out prior to the sample selection. Thus, by considering the sample selection criteria, almost similar or identical enamel characteristics can be selected.

4. Conclusions

Laser irradiation affects dental hard tissues physically and chemically as seen from the percentage difference of elemental compositions and the resulting surface morphology of the samples in this study. This study shows that the irradiation of 100 J/cm^2 obtained less damage on enamel surface compared to the results of higher fluence irradiation. In general, the percentage of carbon increased as an effect of the ablation process, whereas the percentage of oxygen decreased. Phosphorus is not significantly

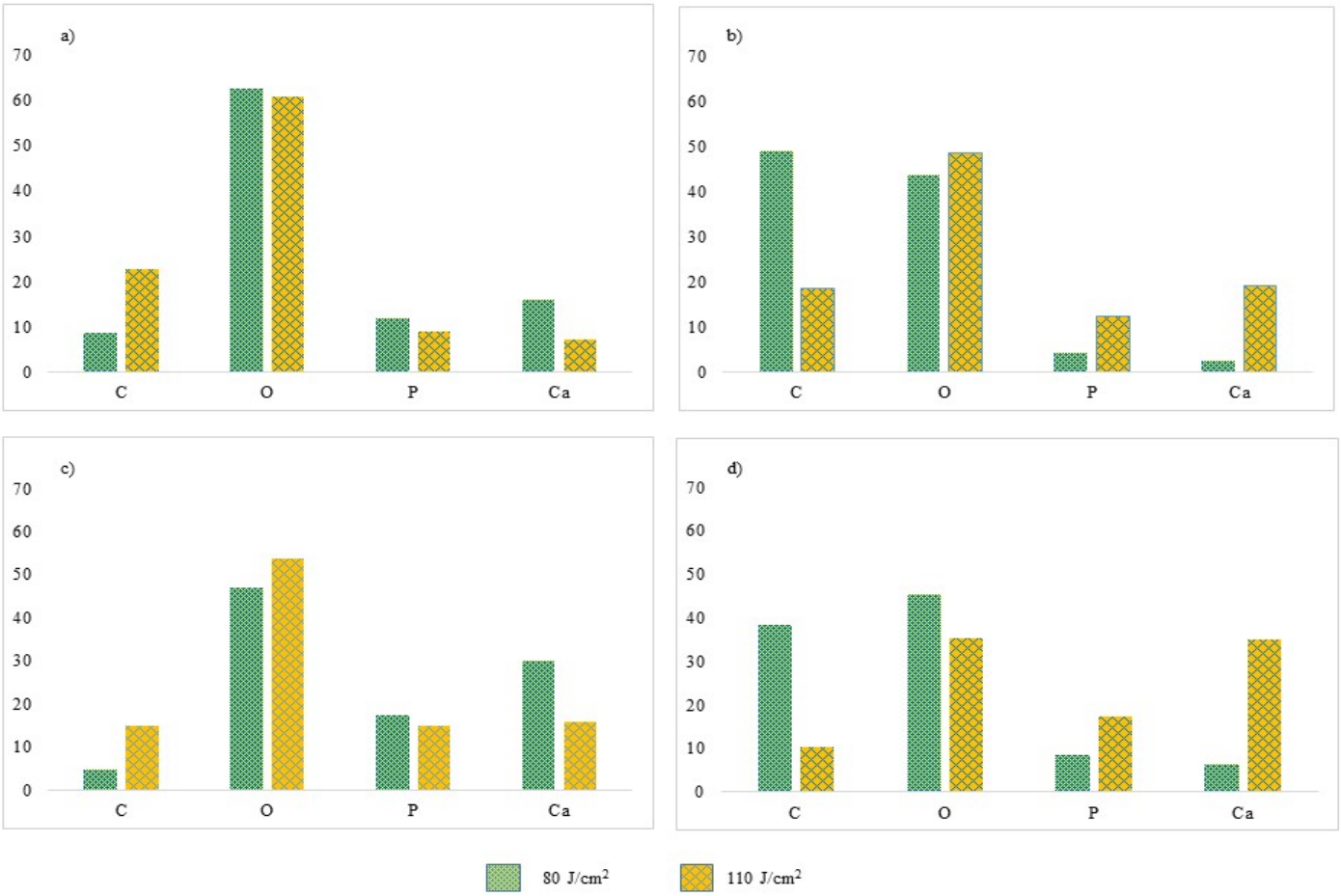


Figure 9: Bar charts of C, O, P and Ca before laser irradiation in terms of a) atomic percentage and b) weight percentage. Bar charts of C, O, P and Ca after laser irradiation of 80, and 110 J/cm^2 at wet condition in terms of c) atomic percentage and d) weight percentage.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to express a profound gratitude to the staffs of Craniofacial and Biomaterial Sciences Cluster, Advanced Medical and Dental Institute and School of Physics, Universiti Sains Malaysia for the support in accomplishing this study. This study is supported by the Short Term Grant from Universiti Sains Malaysia.

References

[1] S. K. Verma, S. Maheshwari, R. K. Singh, P. K. Chaudhari, "Laser in dentistry: An innovative tool in modern dental practice" Natl J Maxillofac Surg, 3(2): 124-132, 2012. https://doi.org/10.4103/0975-5950.111342

[2] M. Migliario, M. Rizzi, A. G. Lucchina, F. Renò, "Diode laser clinical efficacy and mini-invasivity in surgical exposure of impacted teeth" J Craniofac Surg., 27(8), e779-e784, 2016. https://doi.org/10.1097/SCS.0000000000003128

[3] Y. Rezaei, H. Bagheri, M. Esmaeilzadeh, "Effects of laser irradiation on caries prevention" J Lasers Med Sci, 2(4):159-64, 2011. http://dx.doi.org/10.22037/2010.v2i4.2710

[4] S. Kasraei, L. Rezaei-Soufi, E. Yarmohamadi, A. Shabani, "Effect of CO_2 and Nd:YAG lasers on shear bond strength of resin cement to zirconia ceramic" J Dent (Tehran), 12(9), 686–694, 2015.

[5] J. Gan, S. Liu, L. Zhou, Y. Wang, J. Guo, C. Huang, "Effect of Nd:YAG laser irradiation pretreatment on the long-term bond strength of etch-and-rinse adhesive to dentin" Oper Dent., 42(1), 62-72, 2017. https://doi.org/10.2341/15-268-L

[6] N. Gutknecht, "State of the Art in Lasers for Dentistry" J. Laser Health Academy, 2008(3), 1-5, 2008.

[7] A. Beketova, N. Poulakis, A. Bakopoulou, T. Zorba, L. Papadopoulou, D. Christofilos, N. Kantiranis, G. A. Zachariadis, E. Kontonasaki, G. A. Kourouklis, at al., "Inducing bioactivity of dental ceramic/bioactive glass composites by Nd:YAG laser" Dent Mater., 32(11), e284-e296, 2016. 10.1016/j.dental.2016.09.029\

[8] A. Braun, R. F. Krillke, M. Frentzen, C. Bourauel, H. Stark, F. Schelle, "Heat generation caused by ablation of dental hard tissues with an ultrashort pulse laser (USPL) system" Lasers Med Sci, 30(2), 475–481, 2015. https://doi.org/10.1007/s10103-013-1344-z

[9] A. Braun, R. J. Wehry, O. Brede, C. Dehn, M. Frentzen, F. Schelle "Heat generation caused by ablation of restorative materials with an ultrashort pulse laser (UPSL) system" Lasers Med Sci 27(2), 297-303, 2012. https://doi.org/10.1007/s10103-010-0875-9

[10] L. Kuščer, J. Diaci, "Measurement of ebium laser-ablation efficiency n hard dental tissues under different water cooling conditions " J Biomed Opt., 18(10), 108002, 2013. https://doi.org/10.1117/1.JBO.18.10.108002

[11] D. Fried, J. Ragadio, M. Akrivou, J. D. Featherstone, M.W. Murray, K. M. Dickenson, "Dental hard tissue modification and removal using sealed transverse excited atmospheric-pressure lasers operating at lambda=9.6 and 0.6 microm" J Biomed Opt, 6(2), 231-238, 2001. https://doi.org/10.1117/1.1344192
[12] R. C. Cecchini, D. M. Zezell, E. de Oliveira, P. M. de Freitas, P. Eduardo Cde, "Effect of Er:YAG laser on enamel acid resistance: morphlogical and atomic spectrometry analysis." Lasers Surg Med, 37(5), 366-72, 2005. https://doi.org/10.1002/lsm.20247
[13] R. Khatavkar, V. Hegde, "Surface analysis of Erbium:YAG laser etching v/s acid etched surface ESEM observations in vitro study" Laser, 2(1), 2010.
[14] S. Al-Jedani, Y. Al-Hadeethi, M. S. Ansari, M. A. N. Razvi, "Dental hard tisssue ablation with laser irradiation" Austin Dental Sciences, 1(1), 1-10, 2016
[15] Y. Al-Hadeethi, S. Al-Jedani, M. A. Razvi, A. Saeed, A. M. Abdel-Daiem, M. S. Ansari, S. S. Babkair, N. A. Salah, A. Al-Mujtaba, "Data fitting to study ablated hard dental tissues by nanosecond laser irradiation" PLoS One, 11(5), e0156093, 2016. https://doi.org/10.1371/journal.pone.0156093
[16] L. E. Rodríguez-Vilchis, R. Contreras-Bulnes, I. Sánchez-Flores, E. C. Samano, "Acid resistance and structural changes of human dental enamel treated with Er:YAG laser" Photomed Laser Surg., 28(2), 207-11, 2010. https://doi.org/10.1089/pho.2008.2454
[17] L. Bartoli, P. Pouli, C. Fotakis, S. Siano, R. Salimbeni, "Charaerization of stone cleaning by Nd:YAG lasers with different pulse duration" Laser Chemistry, 2006 (8175), 1-6, 2006. https://doi.org/10.1155/2006/81750
[18] I. A. Watson, R. K. Wang, I. Peden, G. D. Ward, D. E. S. Stewart-Tull, A. C. Wardlaw, "Effect of laser and environmental parameters on reducing microbial contamination of stainless steel surfaces with Nd:YAG laser irradiation" Journal of Applied Microbiology, 99, 934-44, 205. https://doi.org/10.1111/j.1365-2672.2005.02665.x
[19] F. M. Suhaimi, N. Z. Zainol Alam, S. Mat Ariffin, N. A. Abd. Razak, M. K. A. A. Razab, "Surface modifications of human tooth using Nd:YAG laser for dental applications" 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Seogwipo, 2017, 4537-4540. https://doi: 10.1109/EMBC.2017.8037865
[20] W. Raucci Neto, C. P. Lepri, J. J. Faraoni Romano, F. S. Fernandes, de L. M. Castro Raucci, L. Bachmann, R. G. Dibb, "Chemical and Morphological Changes of Primary Teeth Irradiated with Nd:YAG Laser: An Ex Vivo Long-Term Analysis" Photomed Laser Surg., 33(5), 266-73, 2015. https://doi.org/10.1089/pho.2014.3876
[21] M. Sydney-Zax, I. Mayer, D. Deutsch, "Carbonate content in developing human and bovine enamel" J Dent Res., 70(5), 913-6, 1991. https://doi.org/10.1177/00220345910700051001
[22] M. F. Little, F. Brudevold, "A study of the inorganic carbon dioxide in intact human enamel" J Dent Res., 37(6), 991-1000, 1958. https://doi.org/10.1177/00220345580370062001
[23] H. W. Kang, I. Rizoiu, A. J. Welch, "Hard tissue ablation with a spray-assisted mid-IR laser" Phys Med Biol, 52(24), 7243-59, 2007. https://doi.org/10.1088/0031-9155/52/24/004

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 407-420 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

An Approach for Determining Rules used to Select Viable Junction Design Alternatives Based on Multiple Objectives

Erwin Bezembinder[*,1], Luc Wismans[2], Eric van Berkum[2]

[1] *Windesheim University of Applied Sciences, Zwolle, The Netherlands*

[2] *University of Twente, Enschede, The Netherlands*

ARTICLEINFO

Article history:
Received: 10 September, 2018
Accepted: 06 October, 2018
Online: 20 October, 2018

Keywords:
Junction design
Junction design assessment
Design rules
Decision tree
CRT
Multiple objectives

ABSTRACT

Transport planners and engineers frequently face the challenge to determine the best design for a specific junction. Many road design manuals provide guidelines for the design and evaluation of different junction alternatives, however these mostly refer to specialized software in which the performances of design alternatives can be modelled. In the first stage of the design process, such assessments of many alternatives are undesirable due to time and budget constraints. There is a need for quick design rules which need limited input data. Although some of these rules exist, their usability is limited due to inconsistencies in rules and non-transparency in combination with objectives. In this paper, we present an approach by which consistent and transparent junction design rules can be determined. The resulting rules can be used to predict a set of viable junction design alternatives for the first stage of the junction design assessment process. The predicted set is in fact the Pareto optimal set of solutions for multiple objectives, e.g. regarding operational, safety and/or environmental impact. The Pareto optimal set of solutions always contains the best solution, whatever set of weights is used for different objectives in a later stage of the assessment process, thus handling multiple objectives in a straightforward manner. The rules are derived from a dataset by using decision tree data mining techniques. For this, a large dataset is first generated, using performance models, with Pareto optimal sets of junction design alternatives for a large amount of, randomly generated, traffic volumes. The approach is applied and evaluated on cases for two different countries. Results show that for over 90% of the situations the Pareto optimal set can be predicted by the new rules, whereas existing rules hardly reach 33%. The new rules provide junction design alternatives with a better performances.

1 Introduction

This paper is an extension of work originally presented in 2017 IEEE 20th International Conference on Intelligent Transport Systems [1]. Where our previous work was a first evaluation of using traditional decision tree algorithms in order to predict a set of (Pareto optimal) junction design alternatives, we now extended the approach, applied and evaluated it on cases for two different countries and compared the results with existing junction design rules by comparing both accuracy and performance.

Junction design involves both the choice for the main type, such as signalised junction or roundabout, as well as choices concerning the number and configuration of the approach and exit lanes, priority handling, slow traffic crossing facilities, signal control and geometric attributes such as the lane length and the central reservation width. The need to determine the best junction design is not only triggered by the construction of new infrastructure. Due to changes in traffic volumes, travel routes and vehicle types there is a regular need for re-evaluation of the junction design.

[*] Erwin Bezembinder, Windesheim University of Applied Sciences, +31 884698436 & e.bezembinder@windesheim.nl

https://dx.doi.org/10.25046/aj030547

The junction design assessment process, which is used to select the best alternative, typically involves three stages [2]. In the first stage, viable alternatives are identified based on limited input data, such as the average annual daily traffic volume, using decision rules, in the form of trees, look-up tables or simple calculation methods, provided in design manuals. In the second stage, more detailed input data, such as the peak hour traffic volumes for each turning movement and specific geometric and control attributes, are collected and the operational, safety and environmental performances are determined by using tools ranging from static analytical to dynamic simulation models. In the third stage, the best alternative is selected based on multiple criteria. The performance measures determined in the previous stage are analysed and weighted in combination with other criteria such as specific local constraints and cost, before selecting the best alternative. This three-stage approach is used to avoid time and money consuming assessment of many alternatives, but possesses the risk of omitting the best solution due to a deficient identification of viable alternatives in the first stage of the assessment process [3]. This deficiency is mostly caused by a lack of consciousness concerning which objectives are served by the rules used in this stage.

The availability and quality of the decision rules used to identify the viable alternatives varies by state, region or country. The rules have been developed over a period of many years based on a combination of data collection and expert judgement. Obviously, this is a positive feature, but also comprises some disadvantages. First, the decision rules do not offer a complete and consistent approach for all junction design alternatives. Some rules are only meant for signalised junctions where others only encompass whether a left-turn bay is needed on a major road approach to a priority junction. Some rules are formal warrants where others are informal guidelines. Second, many decision rules have not been updated for many years and thus do not reckon with changes in traffic behaviour and vehicle types and do not include relatively new junction design alternatives. Third, rules are often based on one underlying objective, generally referring to the operational performance, causing junction design alternatives that are better or best for other objectives to be neglected. Other rules implicitly contain a weighing or preferred order for different objectives, which restricts and complicates the assessment of multiple objectives in a later stage of the junction design assessment process.

Surprisingly, only limited efforts have been made to develop new junction design rules. [4] used the HCM 2010 methodologies [5] to distinguish between different junction types based on the major and minor street volumes. They generated a dataset with more than 6,000 scenarios of varying traffic flows for three main junction types. The results were plotted in 2- and 3D diagrams using the major an minor volumes and the average control delay classified by junction type, in order to identify the areas where a junction type has the lowest delay. [6] used the HCM 2000 methodologies [7] to determine so-called shape-grammars for junction type choice based on the total traffic volume and the through traffic share for three and four arm junctions resulting in look-up tables. Although these studies provided interesting approaches, they examine a limited number of junction designs and demand volumes and moreover, they conducted a manual analysis of the generated data. [8] generated a dataset with the operational performance of 1,296 scenarios for different junction design types and traffic patterns using VISSIM, determining the total delay, stop delay and the average number of stops per vehicle. A classification tree method was then used to group the scenarios into as many homogeneous classes as possible. Classes were constructed for capacity shortage as well as delay. Based on these classes one can easily discover the average operational performance for one or more junction design alternatives. Still, the resulting classes do not provide a set of viable junction design alternatives and are only based on the operational objective.

In this paper, we present an approach by which new junction design rules can be determined. The resulting rules can be used to predict a set of viable junction design alternatives for the first stage of the junction design assessment process. The predicted set is in fact the Pareto optimal set of solutions for multiple objectives, e.g. regarding operational, safety and/or environmental impact. The Pareto optimal set of solutions always contains the best solution, whatever set of weights is used for different objectives in a later stage of the assessment process, thus handling multiple objectives in a straightforward manner. The rules are derived from a dataset by using decision tree data mining techniques. For this, a large dataset is needed, which is first generated, using performance models, with Pareto optimal sets of junction design alternatives for a large amount of, randomly generated, traffic volumes. The approach is applied and evaluated on cases in two different countries.

In the next section, we will first explain the suggested approach. Subsequently, in section 3, the evaluation framework will be discussed. In section 4 the existing junction design rules and the implementation of the approach for the two cases will be explained. The results are discussed in section 5. The paper ends with conclusions in section 6.

2 Approach

In this section we will explain our approach for determining junction design rules used to select viable junction design alternatives based on multiple objectives. The approach consists of three major steps:

1. Define the scope
2. Generate the dataset
3. Determine the decision tree

In the first step, choices are made concerning the junction design alternatives, the traffic flow ranges and

the objectives to be considered. In the second step, a dataset is generated containing a Pareto optimal set of junction design alternatives for each, randomly generated, traffic flow pattern. Each set is determined after running performance models for each objective and junction design alternative. In the third step, decision tree algorithms are used to derive rules from the dataset. The steps will be explained in more detail in the next paragraphs.

2.1 Define the scope

In the first major step, choices are made concerning the:

- Junction designs
- Traffic flows
- Objectives
- Performance models

First, the junction design alternatives to be considered should be defined. This is a list of all the possible junction design alternatives that should be evaluated. Junction design involves both the choice for the main type, such as signalised junction or roundabout, as well as choices concerning the number and configuration of the approach and exit lanes, priority handling, slow traffic crossing facilities, signal control and geometric attributes such as the lane length and the central reservation width. The level of detail depends upon the requirements for the specific case. An important issue involves the necessity to classify the junction design alternatives by a size category. This attribute was introduced in order to prevent large (or expensive) junctions always to be the preferred design, regardless of the traffic flow volumes on the junction.

Second, it should be decided for which traffic flow ranges the junction design alternatives should be evaluated. The traffic flow at least involves the turning volumes for the motorised vehicles on the junction. Additionally, pedestrian and bicycle volumes can be used as well. Typically, peak hour volumes are used, since these are major input variables for the performance model(s) used. The range, i.e. the total traffic volume on the junction, is case specific and dependent upon the chosen junction design alternatives.

Third, it should be decided based on which objectives the junction design alternatives should be evaluated. Typically, minimising the operational performance, e.g. the average delay, is used. Additionally, safety, environmental and financial performance measures can be used. The choice is strongly influences by the possibilities to model the performance measures. The suggested approach assumes at least two objectives in order to determine Pareto optimal sets of junction design alternatives.

Fourth, in close connection with the other choices, it should be decided which performance model will be used. The model uses the traffic flows and junction design alternative as input. There is a wide range of models available to determine the performance for a specific objective, ranging from static analytical to dynamic simulation models. The choice depends upon the models and underlying methodologies that are accepted in a specific country, and/or the accepted level of detail of input data or the calculation time of the models. Separate models can be used for separate objectives.

In order to illustrate our approach, we will use a simplified example case. Table 1 shows six four-arm junction design alternatives, with four main junction types, being all-way stop controlled junction (AW), two-way stop controlled junction (TW), signalised junction (SIG) and roundabout (RA). Two size categories (Sc) are differentiated and the lane number and configuration for the approach and exit lanes of the major and minor roads are defined for each alternative. The lane configuration ↑ means that there is one lane which is shared by left turning, straight and right turning traffic, whereas ↰↑ means that there are two lanes; a separate lane for left turning traffic and a shared lane for straight and right turning traffic.

Table 1: Simplified example with the definition of six junction design alternatives

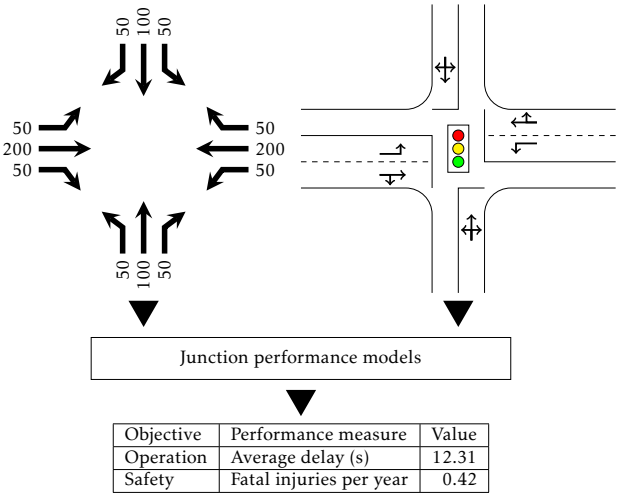
Nr	Id	Type	Sc	Major road		Minor road	
				Approach	Exit	Approach	Exit
1	AW1	AW	1	1 ↑	1	1 ↑	1
2	TW1	TW	1	1 ↑	1	1 ↑	1
3	SIG1	SIG	1	1 ↑	1	1 ↑	1
4	TW2	TW	2	2 ↰↑	1	1 ↑	1
5	SIG2	SIG	2	2 ↰↑	1	1 ↑	1
6	RA2	RA	2	1 ↑	1	1 ↑	1

For the traffic flows, only motorised vehicles are used, where the total traffic flow on the junction is assumed to range between 0 and 2000 pcu/h. Two objectives, being operation and safety will be used, respectively expressed by the (volume weighted) average delay (s) and the estimated number of fatal injuries per year. A fictitious performance model is assumed to generate the measures for this example case. Additionally, a delay threshold of 50 seconds is used, in order to exclude junction design alternatives with an undesirable operational performance.

2.2 Generate the dataset

In the second major step, a dataset is generated that contains a set of junction design alternatives for each combination of traffic demand pattern and size category. This set of junction design alternatives for each combination is the Pareto optimal set of solutions for multiple objectives. The Pareto optimal set is determined by determining the performances for each junction design alternative given a (randomly generated) traffic demand pattern. The performances are determined by using a traffic performance model that uses the traffic demand pattern and junction design as input. It is necessary to generate the data by using performance models, since consistent and comprehensive

data is not available from surveys.



Objective	Performance measure	Value
Operation	Average delay (s)	12.31
Safety	Fatal injuries per year	0.42

Figure 1: Schematic representation of performance modelling for one specific traffic demand pattern and junction design alternative

Figure 1 shows a schematic representation of the performance modelling for one specific traffic demand pattern (with a total volume of 1000 pcu/h) and a junction design alternative (SIG2 from Table 1). This results in a (volume weighted) average delay of 12.31 seconds and an estimated number of fatal injuries of 0.42 per year. For the given traffic demand pattern, this is repeated for each junction design alternative that has been defined. The operational and safety performances of the six junction design alternatives are shown in Figure 2 for both size categories.

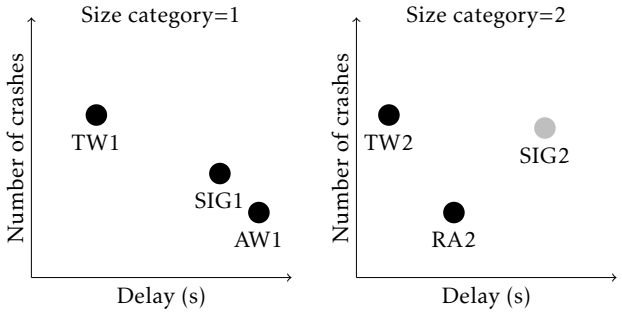


Figure 2: Set of junction design alternative solutions for two different size categories

The main question now is, which junctions design alternatives should be part of the viable set of solutions? The best choice would be the Pareto optimal set of solutions. The Pareto set consists of all solutions for which the corresponding objective values cannot be improved for one objective, without degradation of the other. As said, the Pareto optimal set of solutions always contains the best solution, whatever set of weights are used for the operational or safety performance in a later stage of the assessment process. The Pareto optimal solutions can be determined automatically. In our example, for the first size category, the Pareto optimal solution would be {AW1,TW1,SIG1}. For the second size category it would be {TW2,RA2}. Additionally, solutions with a performance above a certain threshold value, e.g. a specific average delay, could be removed from the viable solution set, in order to prevent analysis of unrealistic solutions in a later stage of the assessment process.

The described process is repeated for multiple (randomly generated) traffic demand patterns. The number of traffic demand patterns to be generated is case specific and should be determined based on the accuracy of the tree to be predicted in the next main step. The accuracy represents how well the tree predicts the Pareto optimal sets that are used as input. This will be explained in more detail in section 3. In earlier experiments [9] we tested the approach for 10-thousand, 100-thousand and 1-million traffic demand patterns. Although, the predictive accuracy of the trees increased with bigger datasets, the differences between the set sizes were minimal. Table 2 shows an example dataset for only four traffic demand patterns. The dataset contains both the variables representing the base traffic pattern, i.e. the volumes for the twelve turning movements as shown in Figure 1 (v1-v12), as well as multiple combined variables such as the total volume on the major (vMa) or minor road (vMi) or the whole junction (vTot). The combined variables are important for the rules to be determined.

Table 2: Simplified example dataset with four traffic demand patterns

Id	v1	v2	...	v12	vMa	vMi	vTot	Sc	Set
1	25	100		25	300	200	500	1	{AW, TW}
2	25	100		25	300	200	500	2	{TW, RA}
3	50	200		50	600	400	1000	1	{AW, TW, SIG}
4	50	200		50	600	400	1000	2	{TW, RA}
5	75	300		75	900	600	1500	1	{SIG}
6	75	300		75	900	600	1500	2	{SIG, RA}
7	100	400		100	1200	800	2000	1	{OTHER}
8	100	400		100	1200	800	2000	2	{RA}

The table also shows a solution set {OTHER}. For this combination of traffic demand pattern and size category, there was no junction design alternative with a delay below the chosen threshold value (50 seconds). The set is included in the dataset in order to facilitate rules that advise an 'other' solution for the specified combination of traffic demand pattern and size category. The created dataset is the foundation of the training set used to determine the decision tree (and the rules) in the next major step.

2.3 Determine the decision tree

In the third major step, the decision tree is determined. Decision trees use a white box model, are easy to understand and interpret, perform well on large datasets, are robust and offer possibilities to validate the model using statistical techniques [10]. Most important, rules can be read from the resulting trees. Determining a tree which can be used to predict a set of junction design alternatives is basically a multi-labelled decision

tree challenge [11]. A modest number of algorithms have been suggested to construct multi-labelled decision trees [12, 13, 14, 15, 16, 17]. In these algorithms, various functions in the traditional decision tree algorithms are replaced by functions fit for handling multi-labelled data, primarily based on measures for the similarity between one or more sets. Unfortunately, these algorithms have some serious disadvantages, the most important one being the fact that they produce very large and complex trees, which is caused by their lack of pruning methods. To overcome these difficulties, we introduced an alternative approach [1]. In this approach we normalise the training data and use the predicted probabilities of the resulting tree, confronted with a threshold value, to determine multiple target labels. This enables us to predict sets of junction design alternatives with traditional algorithms and thus having the advantage of using profoundly proven and widely available methods with a range of modelling options, such as pruning. This approach consists of three (sub)steps:

- Normalise the data
- Built a decision tree
- Determine predicted sets of solutions

2.3.1 Normalise the data

The first step is to apply the standard data normalisation method [18], which is needed when using single labelled decision tree algorithms. This method transforms the data into single-labelled instances. The example training set from Table 2, would then be transformed to the data set with fourteen instances. One instance for each set element. An instance with a set of three junction design alternatives (id 1 in Table 2) is transformed to three instances with one junction design alternative.

2.3.2 Built a decision tree

In the second step, traditional, single-labelled, decision tree algorithms such as ID3, C4.5, CRT, CHAID and QUEST can be used to build a tree. Most of these decision tree algorithms consist of two conceptual phases: growing and pruning. In the tree growing phase, a tree is constructed by using an iterative procedure, were in each iteration the algorithm considers the partition of the training set by selecting the most appropriate attribute according to some splitting measure(s). The tree growing phase continues until a certain stopping criterion is triggered, e.g. when all instances in nodes belong to one class or when a maximum tree depth has been reached. In the pruning phase, the size of the tree is reduced by removing sections of the tree that provide little power to classify the instances. Pruning reduces the complexity of the final tree and hence improves predictive accuracy by reduction of over-fitting.

A major point of criticism for single-labelled decision tree algorithms, is that the obtained classifier could only predict a single label [13]. However, this is only partly true. Although the decision tree algorithms aim to produce nodes and leaves which contain elements with the same label, leaves more often contain elements with multiple labels. This is particularly true for situations where all the instances in a node have (nearly)the same attributes, as a result of which no attribute and value can be determined to split the instances any further. In our example, the normalised instances coming from one set have the same attributes. When there is a sufficient amount of instances in a leaf, i.e. small tree and large dataset, a leaf contains a probability vector, were the probability is determined by comparing the frequency of the instances with a specific label with the total number of instances in the leaf. This probability vector in combination with the normalisation method, offers the opportunity to use the single labelled methods for multiple labelled data.

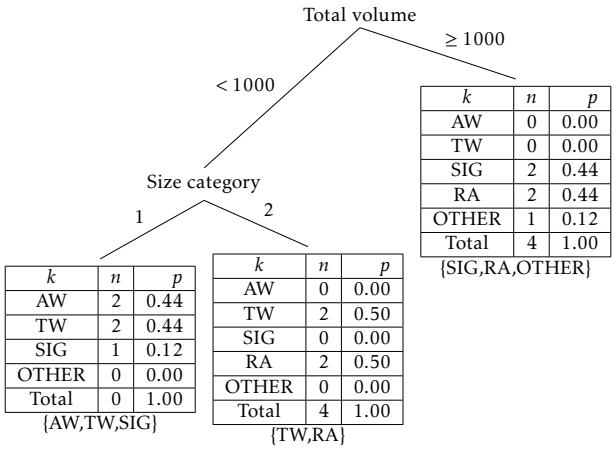


Figure 3: Example decision tree with predicted probabilities

Figure 3 shows the decision tree based on our example training set, but now including the number of instances (count and percentage) for each junction design alternative (k) in each leaf. The percentage represents the predicted probability. Based on preliminary tests we concluded that the CRT decision tree algorithm performed best for our problem. CRT [19] performs binary splits based on a Gini impurity measure and allows pruning.

2.3.3 Determine predicted set of solutions

In the third step, the predicted set of junction design alternatives can be derived from the frequencies and/or predicted probabilities. In our case, a junction design alternative is included in the predicted set if the predicted probability in the leaf is greater than zero. For our example, the predicted set of junction design alternatives is shown at the bottom of each leaf, as can be seen in Figure 3. Alternative threshold values (τ) can be used to determine the predicted sets to reduce the number of junction design alternatives in the sets. Higher values of τ will produce smaller predicted sets of solutions, but with a higher change of missing

junction design alternatives. Lower values of τ will produce larger predicted sets of solutions, but with a higher change of overestimating the number of viable junction design alternatives. The optimal value for τ is to be determined, but can be chosen after the decision tree has been built. Moreover, this variable can either be determined static (1) or dynamic (2):

$$\tau_{static} = a \tag{1}$$

$$\tau_{dynamic} = \frac{1}{(K+b)} \tag{2}$$

Where K is the number of possibly predicted junction design alternatives in the regarding leaf and a and b are constants. In our case K, is dependent upon the size category, since not each size category has the same number of junction design alternatives. In our small example (Figure 3), size category 1 does not include the roundabout, so K would be 4, as can be seen in the leftmost leaf. For size category 2 K is equal to 5, as can be seen in the central leaf. If the size category is not a split attribute, or multiple size categories are combined in one leaf, the highest number for K is presumed, as is the case for the rightmost leaf. Based on preliminary tests we concluded that a dynamic determination of τ with $d = 1$ gave best results.

3 Evaluation framework

We will apply the approach on two cases. The cases will be explained in section 4. In order to determine the success of the approach for the two cases, the new rules resulting from the approach are compared with existing rules. First the predictive accuracy of the rules will be evaluated. This will give insight in how good the rules can reproduce the Pareto optimal set of solutions from the generated dataset as described in section 2.2. Secondly, the performances of the predicted sets of viable junction design alternatives will be evaluated. This will give insight in the average and/or minimal (e.g. operational, safety and environmental) performances of the junction design alternatives that are predicted by the rules. Two-third of the dataset, as described in section 2.2, is used to train the model for the new rules. One-third is used to evaluate both the new and existing rules. The latter is the test set.

3.1 Accuracy

In situations with single-labelled data, the predictive accuracy of a rule is simply the proportion of correctly classified or predicted instances. If we would apply this to our multi-labelled data, this is a very strict measure, being the proportion of instances for which the predicted set of labels is the same as the true set of labels. In our context, predicting a set of viable junction design alternatives for further research, we are also interested in the proportion of instances for which the predicted set at least contains the true set of labels. In this context predicting too much labels instead of predicting too few is less bad. Therefore, we introduce using multiple measures to evaluate the accuracy of the rules. All accuracy measures discussed are on a per instance basis and the aggregate value is an average over all instances. Let T be the true set of labels (i.e. the Pareto optimal set) and P be the predicted set of labels (i.e. by either the new or existing rules). The measure for the predicted set at least containing the true set, we name sufficiency, which is defined as:

$$Sufficiency = \begin{cases} 0, & \text{if } P \not\subseteq T; \\ 1, & \text{if } P \subseteq T. \end{cases} \tag{3}$$

The measure for the predicted set being equal to the true set, we name equality, which is defined as:

$$Equality = \begin{cases} 0, & \text{if } P \neq T; \\ 1, & \text{if } P = T. \end{cases} \tag{4}$$

Although predicting to much labels is less bad than predicting too few, it is still a disadvantage of the model. Therefore, we also evaluate the number of labels that is wrongfully predicted with the overestimation measure:

$$Overestimation = |P/T| \tag{5}$$

This measure is best evaluated in relation with the average set size.

Additionally, we use a similarity measure which is a more general measure for comparing sets used by various authors [13, 14, 17], also referred to as the Jaccard index or intersection over union:

$$Similarity = \frac{|T \cap P|}{|T \cup P|} \tag{6}$$

The values for sufficiency, equality and similarity range between 0 and 1. Higher values correspond to higher accuracy. The value of overestimation ranges from 0 up to the maximum number of labels. Lower values correspond to higher accuracy.

In order to determine the strengths and weaknesses of the new rules, it is imperative to differentiate the indicators for different size and volume categories.

3.2 Performance

An important issue involves the fact that existing rules are often based on one objective, thereby neglecting junction design alternatives that are important for other objectives. If a rule selects viable junction design alternatives based on minimizing the average delay, the alternative with the lowest number of accidents could be missing from the predicted set. It is important to compare the performances resulting from the new and existing rules in relation to the performances resulting from the Pareto optimal set. Various indicators exist for this purpose. [20] provided an excellent overview in the context of comparing two Pareto sets. Here we use the minimum objective value attained by a set for every objective function. This is a straightforward indicator which uncovers the problem mentioned above. This is done by determining the performances for each

junction design alternative in the predicted set. The performances for each objective are determined by using the same junction performance models used to generate the dataset, as described in section 2.2. For each objective, an average minimum performance $\overline{z}_{min}$ is determined by using:

$$\overline{z}_{min} = \frac{\sum_{i=1}^{n} \min\{z_{i1}, z_{i2}, \ldots, z_{ik}\}}{n} \quad (7)$$

where $z_i k$ is the performance of a junction design alternative from the predicted set, given a combination of traffic demand pattern and size category i and n is the total number of these combinations to be evaluated. Using the rules, each combination i produces a predicted set of junction design alternatives. Each of these alternatives has a specific performance for the regarding objective. For each combination of i the minimum performance is used. These values are summed for all combinations of i and divided by the total number of combinations n. This is repeated for each objective and for both the Pareto optimal sets and the sets based on the existing and new rules.

A minor complication is the fact that sets can also contain the 'OTHER' junction design alternative. As explained in section 2.2, for this combination of traffic demand pattern and size category, there was no junction design alternative with a delay below the chosen threshold value. Evidently, since the junction design is unknown, no performances can be determined for this alternative. Therefore, if a set contains an 'OTHER' alternative, the regarding combination of traffic demand pattern and size category i is excluded from equation 7.

4 Cases

Junction design rules can be found in design manuals, which are usually issued by government agencies on a provincial, state or national level. For example, junction design rules are provided in design manuals for the following countries: Australia [21], Belgium [22], Germany [23, 24, 25], France [26], The Netherlands [27, 28, 29], UK [30] and USA [31, 32, 7, 2, 5, 33]. The availability and content of the rules is country specific. This is caused by the fact that policies, procedures, junction design alternatives, traffic modes, traffic behaviour and accepted modelling tools are very different by country. For example, in some countries turbo-roundabouts are generally accepted, whereas in other countries they are not (yet) considered as a viable alternative. Another example is that the same roundabout design can have a different operational and safety performance in different countries. It is not useful to determine universal junction design rules, but rather determine them for a specific region or country.

We chose two case countries, being the United States of America (USA) and the Netherlands, mainly due to the availability of documented junction design rules and corresponding performance models. For both cases, first a short description of the existing rules is given, followed by a brief explanation of the existing rules that haven been chosen for comparison and the application of our approach for the specific case.

4.1 Case 1: United States of America

In this section we will explain the application of our approach as part of a junction design assessment process in the USA.

4.1.1 Existing rules

In the USA, junction design rules can be found in both national and state documents. Here, we will focus on the national level. Junction design rules are scattered over various junction design manuals and technical reports, published by TRB [2, 33], FHWA [34, 32] and AASHTO [31, 35]. A lot of rules require data generated by performance models and lie outside our research scope which aims to determine rules for the first stage of the junction design assessment process. The Manual on Uniform Traffic Control Devices (MUTD) [32] describes eight warrants that define conditions in which a traffic control signal is likely to improve junction safety, operations, or both. If a warrant is met, then signal control may be appropriate. Some warrants require volume based input data, such as eight, four or one hour volumes, while others require performance based data such as delay or the crash history or specific local data such as the area population. Various rules exist for the realisation of specific junction design alternatives and elements, such as volume thresholds for roundabouts or for the number of entry lanes required, e.g. [33]. An overall approach is also provided, combining the earlier mentioned warrants and various other rules in one method [2]. Not all warrants and rules are relevant and/or usable in order to determine a viable set of junction design alternatives. The rules had to be adapted to be used in this research, in order to make a fair comparison between existing and new rules. An extensive description of the existing rules that we used can be found in the Appendix.

4.1.2 Application of the approach

New junction design rules will be determined by using the approach as described in section 2. An important first step is to define the scope, where choices have to be made concerning the junction design alternatives, the traffic flow ranges, the objectives and the performance models to be considered. Table 3 shows the scope definition for this case.

The junction design attributes have been chosen to match the junction design types for the existing rules. Four main types are differentiated. The alternatives are based on the design alternatives mentioned in the underlying existing rules. When there are multiple approach lanes, different lane configurations are possible for the major and minor roads. Multiple variants are then differentiated, leading to 67 separate alternatives to be tested. Seven size categories are differentiated,

which is necessary for the approach to avoid that only the largest junction design types are part of the Pareto optimal set and thus the rules to be determined. An overview of the junction design alternatives can be found in 10 in the Appendix.

Table 3: Scope definition for case 1: USA

Topic	Item	Choices
Junction design	Arms	4
	Main types	AWSC, TWSC, signalised, roundabout
	Alternatives	21
	Variants	67
	Size categories	7
	Attributes	Junction: Main type; Arm: Approach lane configuration, central reservation width, number of exit lanes, sign type, circulating lanes
Traffic flow	Units	Peak hour volumes (pcu/h)
	Modes	Motorised vehicles
	Flows	12 turning flows for motorized vehicles
	Ranges	Total volume: 0-7000 (pcu/h),
Objectives	Objectives	Operation, safety
	Measures	Volume weighted average delay (s), number of fatal-and-injury crashes per year
	Thresholds	Average delay $\leq$ 50 (s)
Models	Operation	HCM 2010 methodologies
	Safety	HSM 2010 methodologies

Random traffic flow patterns are generated based on a total motorised volume ranging between 1 and 7000 (pcu/h). Two objectives and corresponding performance measures have been defined. Junction design alternatives with an average delay above 50 seconds are excluded from the choice sets.

The models used to determine the performances are static analytical based. The operational performance model is an implementation of the Highway Capacity Manual 2010 (HCM2010) methodologies [5]. For signalised junctions control settings (cycle time, green times) are determined (based on the traffic flow and junction design) which aim to minimise the average delay on the junction. The safety performance model is an implementation of the Highway Safety Manual 2010 (HSM2010) methodologies [35].

4.2 Case 2: The Netherlands

In this section we will explain the application of our approach as part of a junction design assessment process in the Netherlands.

4.2.1 Existing rules

In the Netherlands, junction design rules can be found in various publications of the CROW [36, 27, 28, 29]. Most rules are capacity thresholds, which means that they provide a value for the maximum allowed traffic volumes for a certain time interval, e.g. the total traffic volume on a junction or arm in a peak hour (pcu/h). Various tables are provided for various junction design types. Another, rather general but common rule, prescribes a preferred order in which main junction design types should be considered [28, 29]. The preferred order, being; roundabout, priority junction and signalised junction is based on the sustainable safety principle, in which roundabouts are considered (and proven to be) the safest junction design types. The rule dictates that a roundabout should be considered, unless it should be excluded for operational reasons. Furthermore, various design rules exist for specific design elements and/or junction design types. For example, a graphically represented rule which can be used to determine whether a single lane roundabout should be considered for a given combination of entry flow, conflicting circulating flow and crossing bicycle flow [36].

There is no overall (set of) junction design rule(s) available applicable for all relevant junction design alternatives and traffic modes. Therefore, we use a combination of rules from two source. The first source [27] provides two capacity thresholds for different types of junctions,ranging from single lane roundabouts, two lane roundabouts to turbo-roundabouts, priority junctions and signalised junctions. The first threshold represents the maximum allowed sum of all entry flows in a peak hour. The second threshold represents the sum of the entry flow and the (potentially) conflicting flow for one arm and is thus evaluated for each arm. If one of the threshold values is exceeded, then this junction design type is considered unsuitable to be taken into account for further research. The second source [29] provides the so-called Slop-criterion which calculates a value based on the (eight-hour) total volume on the major road, the (eight-hour) maximum volume of the minor roads and a set of parameters. If the resulting value exceeds a certain threshold value, a priority junction is no longer a viable solution. Using the (capacity) thresholds does not provide a set of viable junction design alternatives. However, the rule can be used to determine such a set. An extensive description of the existing rules that we used can be found in the Appendix.

Since bicycle flows are a substantial part of traffic flow in the Netherlands, it is desirable to reckon with these in the rules. However, since there are no existing rules for all junction design types, crossing bicycle flows will be reckoned with in the existing rules by simply adding them to the conflicting flow (1 bicycle = 0.2 pcu), which is used in various rules.

4.2.2 Application of the approach

New junction design rules will be determined by using the approach as described in section 2. An important first step is to define the scope, where choices have to be made concerning the junction design alternatives, the traffic flow ranges and the objectives to be considered. Table 4 shows the scope definition for this case.

The junction design attributes have been chosen to match the junction design types for the existing rules. Four main types and 16 junction designs are differentiated. When there are multiple approach lanes, different lane configurations are possible for the major and minor roads. Multiple variants are then differentiated, leading to 42 separate alternatives to be tested. Five size categories are differentiated, which is neces-

sary for the approach to avoid that only the largest junction design types are part of the Pareto optimal set and thus the rules to be determined. An overview of all junction design alternatives and variants can be found in the Appendix in Table 17.

Table 4: Scope definition for case 2: The Netherlands

Topic	Attribute	Choices
Junction design	Arms	4
	Main types	Equal, priority, signalised, roundabout
	Alternatives	16
	Alternatives	42
	Size categories	5
	Attributes	Junction: Main type; Arm: Approach lane configuration, central reservation width, number of exit lanes, sign type, number of opposed circulating lanes
Traffic flow	Units	Peak hour volumes (pcu/h)
	Modes	Motorised vehicles, bicycles
	Flows	12 turning flows for motorised vehicles, 4 crossing flows bicycles
	Ranges	Total motorised volume: 1-7000 (pcu/h), total bicycle volume 0-500 (bicycles/h)
Objectives	Objectives	Operation, safety, environment
	Measures	Volume weighted average delay (s), number of fatal-and-injury crashes per year, NO_x/PM_{10} emission (g)
	Thresholds	Average delay ≤ 50 (s)
Models	Operation Safety Environment	Local static analytical models

Random traffic flow patterns are generated based on a total motorised volume ranging between 1 and 7000 (pcu/h) and a total bicycle volume ranging between 0 and 500 (bicycles/h). Three objectives and corresponding performance measures have been defined. Junction design alternatives with an average delay above 70 seconds are excluded from the choice sets.

The models used to determine the performances are static analytical based and were specially developed and tested for use in the Dutch practice [37]. The operational performance model is similar to the methodologies provided in the American HCM2010 [5] and the German HBS2015 [25], extended with new methods for (turbo)roundabouts and the effects of bicycle traffic. The safety performance model is similar to the American HSM2010 [35] methodologies, where parameters were adopted from [38] for local situations. The environmental performance model determines the emissions based on local emission substance factors multiplied by the queue on each lane [39, 3].

5 Results

In this section we will discuss the results. We will first consider the predictive accuracy of the existing and newly determined junction design rules. Next we will evaluate the performances of the predicted sets of viable junction designs. All results were determined with the test dataset.

5.1 Accuracy

Table 5 shows accuracy values for both the existing and new rules for both cases. The accuracy represents how good the rules can predict the Pareto optimal sets. The table shows values for sufficiency, similarity, equality and overestimation, according to equations 3-5. In addition to the accuracy values, the average set size is presented.

Table 5: Accuracy measures of existing and new rules for both cases

Item	Case 1		Case 2	
	Existing	New	Existing	New
Sufficiency	0.311	0.905	0.845	0.899
Similarity	0.249	0.723	0.391	0.821
Equality	0.129	0.492	0.051	0.745
Overestimation	1.320	0.550	1.596	0.293
Average set size	1.857	1.741	2.606	1.242

The table shows that the new rules can better predict the Pareto optimal sets then the existing rules can. For both cases, the sufficiency values, being the measure for the predicted set at least containing the true (Pareto optimal) set, for the new rules are higher. For case 1 the difference is evident. The existing rules have a sufficiency rate of 31.1% where the new rules accomplish a 90.5%. For case 2, the difference in sufficiency rates is limited, with respectively 84.5% and 89.9% for the existing and new rules. However, there are big differences for the other measures. This is caused by the fact that the existing rules predict large sets, with an average set size of 2.606 and an average overestimation of 1.596. Larger sets have a higher change to at least contain the true set, which is reflected in the sufficiency value. However, equality, which reflects the rate at which the predicted set being equal to the true set, is extremely low. The rules are not discriminating enough, thus causing substantially more work in a later stage of the junction design assessment process. The new rules produce more equal and smaller sets.

The presented accuracy values in Table 5 are average values for the whole test set. In order to determine whether differences exist for different size or volume categories, differentiated results are presented in Table 6 and Table 7.

Table 6: Accuracy measures for the new rules differentiated by size category for both cases

Case	Size category	Sufficiency	Similarity	Equality	Overestimation
1	1	0.951	0.826	0.674	0.404
	2	0.899	0.860	0.807	0.182
	3	0.882	0.807	0.586	0.375
	4	0.882	0.747	0.594	0.604
	5	0.873	0.583	0.234	0.805
	6	0.933	0.578	0.216	0.840
	7	0.913	0.660	0.335	0.642
2	1	0.954	0.902	0.851	0.153
	2	0.949	0.844	0.723	0.285
	3	0.894	0.857	0.802	0.179
	4	0.870	0.760	0.678	0.437
	5	0.830	0.743	0.672	0.410

In Table 6, which shows the accuracy measures differentiated by size categories, it can be seen that sufficiency rates are fairly stable for all size categories. The values range between 87.3% and 95.1% for case 1 and between 83.0% and 95.4% for case 2. For case

2, equality rates are also relatively high. For case 1, these rates are lower for the higher size categories. This means that the new rules predict bigger sets, which also can be seen while looking at the overestimation values. This is not a big problem, since the overestimation is still limited and moreover, the similarity values, which represent an average set comparison, are still reasonably high.

In Table 7, which shows the accuracy measures differentiated by volume categories, a similar stable set of sufficiency rates can be seen for both cases.

Table 7: Accuracy measures for the new rules differentiated by volume category for both cases

Case	Volume category (pcu/h)	Sufficiency	Similarity	Equality	Overestimation
1	<1000	0.961	0.791	0.506	0.527
	1000-1999	0.914	0.741	0.484	0.605
	2000-2999	0.711	0.557	0.331	0.797
	3000-3999	0.851	0.681	0.477	0.555
	4000-4999	0.934	0.743	0.533	0.484
	5000-5999	0.970	0.767	0.551	0.451
	≥6000	0.984	0.773	0.557	0.443
2	<1000	0.931	0.826	0.706	0.292
	1000-1999	0.859	0.826	0.768	0.226
	2000-2999	0.857	0.773	0.666	0.315
	3000-3999	0.931	0.668	0.450	0.743
	4000-4999	0.865	0.798	0.763	0.338
	5000-5999	0.906	0.906	0.906	0.094
	≥6000	0.943	0.943	0.943	0.057

Sufficiency rates are best for the lower and higher volume categories. The middle categories generally have the lowest sufficiency values. This can in part be explained by the fact that there are more junction design alternatives available for these volume categories. Moreover, as was the case for the size categories, the similarity values are again still reasonably high.

Generally, it can be concluded that the new rules have a much better accuracy then the existing rules and that there are no specific weaknesses concerning the accuracy values if they are differentiated for size and volume categories. The Pareto optimal sets can be predicted with a 90% accuracy.

5.2 Performance

Table 8 shows the average minimum performances for all objectives as has been defined in equation 7. The operational performance is expressed as volume weighted average delay (s), the safety performance as the number of fatal-and-injury crashes by year, and the environmental performance as the PM_{10} emissions (g). The table shows the average minimum performances for respectively the Pareto optimal sets, the sets predicted by the existing rules and the sets predicted by the new rules. For a fair comparison, if any of the sets (Pareto, existing, new) only contains the 'OTHER' junction design alternative, the regarding instance is excluded from the analysis, as was mentioned in section 3.2.

Table 8: Performances measures of existing and new rules for both cases

$\bar{z}_{min}$	Case 1			Case 2		
	Pareto	Existing	New	Pareto	Existing	New
Operation	16.415	24.476	18.588	15.579	23.995	31.367
Safety	0.787	0.782	0.776	0.581	0.577	0.574
Environment				0.923	1.869	1.028

For case 1, the new rules show that the average minimum performance for both objectives is better (lower) in comparison with the existing rules. This means that the new rules produce sets of junction design alternatives which contain alternatives with a better (lower) performance for either operation, safety or a combination of one or more of these objectives. This was to be expected, since the sufficiency rate for the existing rules was only 31.1% for this case. On the other hand, the differences are not as big as those for the accuracy rates. The new rules provide an average reduction of 24% for operational and only 1% for safety performances in comparison with the existing rules. This is partly caused by the exclusion of instances for which one of the sets predicts a single 'OTHER' junction design alternative. This was the case for 58% of the instances. A striking point is the fact that the average minimum safety performance for both the existing and the new rules is lower than that for the Pareto optimal set. This seems strange, since an alternative that has the minimum (optimal) performance for one objective is to be expected to be in the Pareto optimal set. This was true, however since we used a threshold value for the operational performance, some alternatives were excluded from the eventual set. When the new rules 'accidentally' predicted these alternatives, the performance measures for these alternatives were yet reckoned with.

For case 2, the average minimum operational performance is better for the existing rules than for the new rules. This is primarily caused by the fact that the existing rules predict large sets, as was mentioned in the previous section, thereby increasing the change that the junction design alternative with the minimum performance is part of the predicted set. Moreover, the existing rules are specifically made to serve the operational objective. On the other hand, the average minimum performance for the new rules is rather high. Further analysis shows that there is a relatively small amount of wrongly predicted junction design alternatives which have very high values for the operational performance. Table 9 shows the number of instances for various delay classes.

Table 9: Number of instances by delay class for case 2

Delay class (s)	Pareto	Existing	New
< 25	5389	5224	5227
≥ 25 and < 50	1157	1183	1185
≥ 50 and < 70	459	498	480
≥ 70	0	70	113

When using median values instead of average, the

operational performances for the Pareto, existing and new sets are respectively 7.145, 7.175 and 7.156. The performance values for the safety and environmental performances are better (lower) for the new rules. The new rules provide an average reduction of 1% for the safety performance and 45% for the environmental performance.

Generally, it can be concluded that the new rules predict junction design alternatives with either equal or better (lower) performances for all objectives.

6 Conclusions

In this study, we demonstrated the application of a new approach to determine junction design rules for use in the first stage of the junction design assessment process. The approach uses decision tree algorithms to determine decision trees based on modelled data. There is a need for quick design rules which need limited input data. Although some of these rules exist, their usability is limited due to inconsistencies in rules and non-transparency in combination with objectives. In this paper, we present an approach by which new and better junction design rules can be determined. The resulting rules can be used to predict a set of viable junction design alternatives for the first stage of the junction design assessment process. Moreover, the predicted set is in fact the Pareto optimal set of solutions for multiple objectives, e.g. regarding operational, safety and/or environmental impact. The Pareto optimal set of solutions always contains the best solution, whatever set of weights is used for different objectives in a later stage of the assessment process, thus handling multiple objectives in a straightforward manner.

The approach was applied and evaluated on cases in two different countries. Results show that for about 90% of the situations the Pareto optimal set can be predicted by the new rules, whereas existing rules hardly reach 35% or are not discriminating enough resulting in large set sizes, creating more work in later stages of the junction design assessment process. The new rules provide better performances for the non-operational objectives (safety and environment). Results for the operational performances were different for the two cases. Generally, the new rules produce better results with smaller predicted sets.

The major contribution of this paper is that it presents an approach to determine consistent and complete junction design rules based on modelled data in a transparent and systematic manner. Although some efforts have been made to determine junction design rules based on modelled data no generic and transferable method had been developed until now.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment Time New Roman, 10 Normal. Acknowledge your institute/ funder.

Appendix

In this appendix we provide a description of the junction design alternatives and existing junction design rules for respectively case 1 and 2. At the end of the appendix, an explanation of the variable notations is provided.

Case 1: United States

Table 10 shows the junction design alternatives used for this case.

Table 10: Junction design alternatives for case 1

Nr	Id	Sc	Approach lane number and configuration(s)	
			Major road	Minor road
1	A11	1	1	1
2	T11	1	1	1
3	S11	1	1	1
4	T21	2	2 	1
5	S21	2	2	1
6	T22	3	2	2
7	T31	3	3	1
8	S22	3	2	
9	S31	3	3	1
10	T32	4	3	2
11	S32	4	3	2
12	S41	4	3 	1
13	1R11	4	1	1
14	S33	5	3	3
15	S42	5	4	2
16	S43	6	4	3
17	S44	6	4	4
18	2R11	6	1	1
19	S64	7	6 	4
20	2R21	7	2	1
21	2R22	7	2	2

A=AWSC-junction,T=TWSC-junction,S=signalised junction, 1R=one-lane roundabout,2R=two-lane roundabout

There are 21 junction design alternatives. Each alternative can be identified by the main type and the number of approach lanes on the major and minor road arms. Junction design alternative number 4, with id 'S21' is a signalised junction with two approach lanes on the major road arms and one approach lane on the minor road arms. A description of the characters uses for the junction types is provided at the bottom of the table. The number of exit lanes is automatically determined, based on the number of approach lanes with a destination to the regarding arm. When there are multiple approach lanes, different lane configurations are possible. Multiple variants are then differentiated, e.g. three (3 times 1) variants for alternative number 4 and nine (3 times 3) variants for alternative number 15. In total 67 separate variants are defined. The junction design alternatives are grouped in seven size categories (Sc), based on the required space. The rules are based on the guidelines provided in [2]. We only use guidelines that refer to a specific junction design alternative. Rules referring to more detailed design elements, such as on-street parking, left-

turn prohibition, lane length, right-turn radius and signal flash mode, are excluded. Obviously, we only use guidelines that need volume-based input. Guidelines that require performance-based input, such as delay or crash-history, are excluded, since we don't want to run a model or perform an extensive survey in the first stage of the junction assessment process. The guidelines regarding signalised junctions in [2] are based on the so-called traffic signal warrants from [40]. These warrants have been updated over the years and therefore we use the most recent version from [32]. Most guidelines regarding stop-controlled junctions are presented as graphs. For these, different (linear, exponential, logarithmic, second-order polynomial) functions have been estimated to fit the lines in the graphs.

Table 11 shows which rules are applied for each size category (Sc). Each size category has a base set of viable junction design alternatives, corresponding to the definitions in Table 10. For each alternative in the base set it is evaluated whether it should be included (I) or excluded (E) from the base set. For each main type (A=AWSC-junction, T=TWSC-junction, S=signalised junction and R=roundabout), this is determined by one ore more rules. Each rule performs a check resulting in a true or false value. If true then that particular type is either included or excluded from the base set. Table 11 shows which rules cause inclusion or exclusion. As a starting point all alternatives are excluded. The rules are executed in the order as shown in Table 11. For example, for size category (Sc) 1, 'S11' is excluded unless one of the rules S1-S4 provides a true value. Alternative 'T11' is excluded unless rule T1 is true, but if one of the rules T2-T4 is true it is always excluded. Table 12 provides the conditions that are checked for each rule.

Table 11: Junction design rules for case 1: Application of rules by size category

Sc	Base set	Rules											
		S1	S2	S3	S4	A1	A2	T1	T2	T3	T4	T5	R1
1	{A11,T11,S11}	I	I	I	I	I	E	I	E	E	E		
2	{T21,S21}	I	I	I	I			I	E			E	
3	{T31,S31,T22,S22}	I	I	I	I			I	E			E	
4	{T32,S32,S41,1R11}	I	I	I	I			I					I
5	{S33,S42}	I	I	I	I								
6	{S43,S44,2R11}	I	I	I	I								I
7	{S64,2R21,2R22}	I	I	I	I								I

Table 12: Junction design rules for case 1: Rules and conditions

Code	Condition	Specifics
S1	$q_{mas8} \geq \lambda_{S1a}$ and $q_{mix8} \geq \lambda_{S1b}$	Signal warrant 1a
S2	$q_{mas8} \geq \lambda_{S2a}$ and $q_{mix8} \geq \lambda_{S2b}$	Signal warrant 1b
S3	$q_{mas4} \geq 0.0003 q_{mix4}^2 - \lambda_{S3a} q_{mix4} + \lambda_{S3b}$	Signal warrant 2
S4	$q_{mas1} \geq 0.0002 q_{mix1}^2 - \lambda_{S4a} q_{mix1} + \lambda_{S4b}$	Signal warrant 3
A1	$q_{mas8} \geq 300$	
A2	$J = signal$	
T1	$q_{mas24} \geq 6000$	
T2	$q_{mi1} > \lambda_{T2a} e^{(-0.001 q_{mas1})}$	≥ 2 lanes on minor road
T3	$q_{mao1} > \lambda_{T3a} \ln(q_{ma1}) + \lambda_{T3b}$	left-turn lane on major road
T4	$q_{mar1} > 683.6 e^{(-0.004 q_{mas1})}$	right-turn lane on major road
T5	T3 and T4	
R1	$q_{tot24} \geq 3600 + 9000 l_c (1 + (81/50)) - 94 p_l$	

Tables 13, 14 and 15 provide parameter values for the λ that is used in various rules. Table 16 provides the volume conversion factors that were used to determine the volumes needed for the rules, based on the peak hour volume flows (pcu/h) that are used as base input. The variable notations are explained at the end of the appendix.

Table 13: Junction design rules for case 1: Parameters for rules S1-S4

l_{ma}	l_{mi}	λ_{S1a}	λ_{S1b}	λ_{S2a}	λ_{S2b}	λ_{S3a}	λ_{S3b}	λ_{S4a}	λ_{S4b}
1	1	500	150	750	75	0.7697	573.8	0.7815	751.44
≥2	1	600	150	900	75	0.7689	659.11	0.7339	842.04
≥2	≥2	600	200	900	100	1.0226	881.44	0.9397	1076.4

Table 14: Junction design rules for case 1: Parameters for rule T2

p_{mir}	λ_{T2a}
≤ 0.35	550.67
> 0.35	641.80

Table 15: Junction design rules for case 1: Parameters for rule T3

p_{mal}	λ_{T3a}	λ_{T3b}
< 0.075	-922.4	6157.0
$0.075 - 0.124$	-951.8	6046.6
$0.125 - 0.174$	-863.7	5369.8
$0.175 - 0.299$	-997.9	5849.3
≥ 0.300	-880.7	5362.9

Table 16: Volume conversion factors for the junction design rules of case 1

Volume	Units	Factor
Peak hour volume	pcu/h	1.00
Peak hour volume	veh/h	0.93
Four-hour volume	veh/h	0.78
Eight-hour volume	veh/h	0.67
Daily volume	veh/day	10.33

Case 2: The Netherlands

Table 17 shows the junction design alternatives used for this case. There are 16 junction design alternatives. As for case 1, each alternative can be identified by the main type and the number of approach lanes on the major and minor road arms. The characters referring to the main junction types are slightly different compared to case 1, and are provided at the bottom of the table. In total 42 separate variants are defined. The junction design alternatives are grouped in five size categories (Sc), based on the required space.

The rules are based on guidelines provided in [27] and [29]. There are three main types of rules. The first rule uses a threshold value for the total amount of traffic on the junction. Different threshold values are used for different junction types. The second rule checks a threshold value regarding the conflicting volume for each arm. If the threshold is exceeded for one

of the arms, a specific type is not viable. The third rule is the so-called Slop-criterion, which is used to check whether a priority junction is still viable. Table 18 shows which rules are applied for each size category (Sc).

Table 17: Junction design alternatives for case 2

Nr	Id	Sc	Approach lane number and configuration(s)	
			Major road	Minor road
1	E11	1	1	1
2	P11	1	1	1
3	S11	1	1	1
4	P21	2	2	1
5	P22	2	2	1 2
6	S21	2	2	1
7	S22	2	2	2
8	S33	3	3	3
9	1R11	3	1	1
10	2R11	3	1	1
11	S44	4	4	4
12	TR22	4	2	2
13	2R22	4	2	2
14	S66	5	6	6
15	TR23	5	2	3
16	TR33	5	3	3

E=equal junction,P=priority junction,S=signalised junction,
1R=one-lane roundabout,2R=two-lane roundabout,TR=turboroundabout

Table 18: Junction design rules for case 2: Application of rules by size category

Sc	Base set	Rules									
		E1	E2	E3	P1	P2	P3	S1	S2	R1	R2
1	{E11,P11,S11}	E	E	E	E	E	E	E	E		
2	{P21,P22,S21,S22}						E	E	E		
3	{S33,1R11,2R11}							E	E	E	E
4	{S44,2R22,TR22}							E	E	E	E
5	{S66,TR23,TR33}							E	E	E	E

In this case, as a starting point all alternatives for a specific size category are included in the final set unless one of the applied rules for that specific alternatives returns a true value. Table 19 provides the conditions that are checked for each rule.

Table 19: Junction design rules for case 2: Rules and conditions

Code	Condition	Specifics
E1	$q_{tot1} > 1500$	total volume threshold
E2	$q_c > 1100$	conflicting volume threshold
E3	$(q_{mix8}/\beta_1) \times (-1 + \sqrt{1 + \beta_2 q_{mas8}/q_{mix8}}) > 1.33$	Slop-criterion
P1	$q_{tot1} > 1500$	total volume threshold
P2	$q_c > 1100$	conflicting volume threshold
P3	$(q_{mix8}/\beta_1) \times (-1 + \sqrt{1 + \beta_2 q_{mas8}/q_{mix8}}) > 1.33$	Slop-criterion
S1	$q_{tot1} > \lambda_{S1}$	total volume threshold
S2	$q_c > \lambda_{S2}$	conflicting volume threshold
R1	$q_{tot1} > \lambda_{R1}$	total volume threshold
R2	$q_c > \lambda_{R2}$	conflicting volume threshold

Table 20: Junction design rules for case 2: Parameters for rules E3,P3

l_{ma}	l_{mi}	β_1	β_2
1	1	300	2.0
2	1	300	2.4
1	2	400	3.2
2	2	400	2.7

Tables 20,21 and 22 provide various parameter values for the rules. In this case only one volume conversion value is used in comparison with the peak hour volume (1.00). In order to obtain the eight-hour volume a factor of 0.63 is used.

Table 21: Junction design rules for case 2: Parameters for rules S1-S2

l_{ma}	λ_{S1}	λ_{S2}
< 4	3500	3800
≥ 4	7500	3800

Table 22: Junction design rules for case 2: Parameters for rules R1-R2

Roundabout alternative	λ_{R1}	λ_{R2}
1R11	2000	1500
2R11	2200	1700
2R22	3500	2400
TR22	3500	2100
TR23	4000	2300
TR33	4500	2800

Notations

Variables:

q_{at}	Vehicular traffic volume for turn(s) a and time unit t
λ_r	Parameter for rule r
p_m	Percentage of traffic volume for movement m
l_c	Number of circulating lanes on a roundabout
l_{ma}	Number of lanes for the major approach
l_{mi}	Number of lanes for the minor approach
J	Junction type

Subscripts (a, t, m):

1	Peak hour volume
4	Four-hour volume
8	Eight-hour volume
24	Daily volume
ma	Major arm approach volume
mao	Major arm opposing volume (right-turn and through)
mar	Major arm right-turn volume
mas	Sum of approach volumes of both major arms
mi	Minor arm approach volume
mir	Minor arm turn-turn volume
mix	Maximum approach volume of both minor arms
tot	Sum of approach volumes of all arms
l	Left-turn movement volume
c	Conflicting volume for an arm

References

[1] E. M. Bezembinder, L. J. J. Wismans, and E. C. van Berkum. Constructing multi-labelled decision trees for junction design using the predicted probabilities. In *Proceedings IEEE 20th International Conference on Intelligent Transportation Systems*, Yokohama, Japan, October16-19 2017.

[2] TRB. Evaluating intersection improvements: An engineering study guide. NCHRP Report 457, Transportation Research Board of the National Academies, Washington D.C., United States, 2001.

[3] E. M. Bezembinder, L. J. J. Wismans, and E. C. van Berkum. Multi-objective assessment framework for intersection alternatives under different demands. In *Proc. Transportation Research Board 95th Annual Meeting*, Washington D.C., United States, January10-14 2016.

[4] L.D. Han, J-M Li, and T. Urbanik. Control-type selection at isolated intersections based on control delay under various demand levels. *Transportation Research Record: Journal of the Transportation Research Board*, 20701:109–116, 2008.

[5] TRB. *Highway Capacity Manual 2010*. Washington D.C., United States, 2010.

[6] B. J. Vitins and K. W. Axhausen. Shape grammars for intersection type choice in road network generation. In *Proc. 12th Swiss Transportation Research Conference*, Monte Verita/Ascona, Switzerland, May2-4 2012.

[7] TRB. *Highway Capacity Manual 2000*. Washington D.C., United States, 2000.

[8] A. Tarko, M. S. Azam, and M. Inerowicz. Operational performance of alternative types of intersections - a systematic comparison for indiana conditions. In *Proc. 4th International Symposium on Highway Geometric Design*, Valencia, Spain, June5-9 2010.

[9] E. M. Bezembinder, L. J. J. Wismans, and E. C. van Berkum. Using decision trees in order to determine intersection design rules. In *Proc. Transportation Research Board 94th Annual Meeting*, Washington D.C., United States, January11-15 2015.

[10] L. Rokach and O. Maimon. *Data mining with decision trees*. World Scientific Publising Co. Pre. Ltd., Singapore, 2008.

[11] S. Kotsiantis. Decision trees: a recent overview. *Artificial Intelligence Review*, 39:261–283, 2013.

[12] A. Clare and R. King. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conference on PKDD*, Freiburg, Germany, 2001.

[13] Y.L. Chen, C.L. Hsu, and S.C. Chou. Constructing a multi-valued and multi-labelled decision tree. *Expert Systems with Applications*, 25:199–209, 2003.

[14] S.C. Chou and C.L. Hsu. Mmdt: a multi-valued and multi-labeled decision tree classifier for data mining. *Expert Systems with Applications*, 28:799–812, 2005.

[15] H.Li, R. Zhao, J. Chen, and Y. Chang. *Research on multi-valued and multi-labeled decision trees*, pages 247–254. Springer Berlin Heidelberg, 2006.

[16] W. Yi, M. Lu, and Z. Liu. Multi-valued attribute and multi-labeled data decision tree algorithm. *International Journal of Machine Learning and Cybernetics*, 2:67–74, 2011.

[17] W. Yi, L. Yan, M. Lu, and Z. Liu. A new multi-valued attribute and multi-labeled data decision tree algorithm. *Information*, 16(7A):4771–4780, 2013.

[18] E.F. Codd. *Further normalization of the data base relational model*, volume 6 of *Courant computer science symposium*, pages 33–64. NJ: Prentice-Hall, Englewood Cliffs, 1972.

[19] L. Breiman, J. Friedman, R. Olshen, and C.Stone. *Classification and regression trees*. Wadsworth International Group, Belmont, California, 1984.

[20] T. Brands. *Multi-objective optimisation of multimodal passenger transportation networks*. PhD thesis, University of Twente, Enschede, The Netherlands, 2015.

[21] Austroads. Guide to traffic management part 6: Intersections, interchanges and crossings. Technical report, Austroads, Sydney, Australia, 2013.

[22] AWV. Vadumecum veilige wegen en verkeer. Technical report, Vlaamse Overheid, Agentschap Wegen en Verkeer, Brussels, Belgium, 2009.

[23] FGSV. Richtlinien fr die anlage von stadstrassen: Rast06. Technical Report 200, Forschungsgesellschaft fr Strassen- und Verkehrswesen, Kln, Germany, 2007.

[24] FGSV. Richtlinien fr die anlage von landstrassen: Ral. Technical Report 201, Forschungsgesellschaft fr Strassen- und Verkehrswesen, Kln, Germany, 2013.

[25] FGSV. Handbuch fr die bemessung von strassenverkehrsanlagen. Technical Report 299, Forschungsgesellschaft fr Strassen- und Verkehrswesen, Kln, Germany, 2015.

[26] SETRA. The design of interurban intersections on major roads. Technical report, Service d'tudes Techniques des Routes et Autoroutes, Bagneux Cedex, France, 2002.

[27] CROW. Turborotondes. Technical Report Publication 257, CROW, Ede, The Netherlands, 2008.

[28] CROW. Asvv 2012. Technical Report Publication 723, CROW, Ede, The Netherlands, 2012.

[29] CROW. Handboek wegontwerp 2013. Technical Report Publications 328-331, CROW, Ede, The Netherlands, 2013.

[30] HE. *Design Manual for Roads and Bridges*. London, England, online version march 2017 edition, 2017.

[31] AASHTO. *A Policy on Geometric Design of Highways and Streets, 6th Edition*. Washington D.C., United States, 2010.

[32] FHWA. Manual on uniform traffic control devices for streets and highways (mutcd). Technical Report 2009 Edition with Revisions 1 and 2, May 2012, Federal Highway Agency of the United States Department of Transportation, Washington D.C., United States, 2012.

[33] TRB. Roundabouts: An informational guide. NCHRP Report 672, Transportation Research Board of the National Academies, Washington D.C., United States, 2010.

[34] FHWA. Alternative intersections/interchanges: Informational report (aiir). Technical Report FHWA-HRT-09-060, FHWA, Washington D.C., United States, 2010.

[35] AASHTO. *Highway Safety Manual*. Washington D.C., United States, 2010.

[36] CROW. Eenheid in rotondes. Technical Report Publication 126, CROW, Ede, The Netherlands, 1998.

[37] E. M. Bezembinder. Junction modelling 2.0. Technical report, Windesheim, Zwolle, The Netherlands, 2018.

[38] A. Dijkstra. Enkele aspecten van verkeersveiligheid. Research report R-2014-21A, Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV, Den Haag, The Netherlands, 2014.

[39] R. H. Gasthuis. Emission modelling on junctions in urban networks. Master's thesis, University of Twente, Enschede, The Netherlands, 2015.

[40] FHWA. Manual on uniform traffic control devices for streets and highways (mutcd). Technical report, Federal Highway Agency of the United States Department of Transportation, Washington D.C., United States, 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 5, 427-438 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Improved DTC Control Strategy of B12 Inverter Fed BLDC Motor Drives Considering Commutation Torque Dips

Rabiaa Mars*, Badii Bouzidi, Bassem El Badsi, Abderrazak Yangui

University of Sfax, Electrical Department, Sfax Engineering National School, P.O. Box 1173, 3038 Sfax, Tunisia

A R T I C L E I N F O

Article history:
Received: 10 July, 2018
Accepted: 09 October, 2018
Online: 24 October, 2018

Keywords:
Brushless DC motor
Commutation torque dips
Direct torque control scheme
Current slope
Simulation results

A B S T R A C T

In most industrial high-power applications, the brushless DC (BLDC) motor is connected to multi-level inverters specially the three-level NPC inverters (B12). Torque ripple is a critical issue in BLDC motor drives. Accordingly, minimizing the ripple produced in the torque is necessary to enhance the BLDC motor drive performances. This paper aims to develop two direct torque control (DTC) schemes. To evaluate the proposed strategies performances, three DTC schemes, namely DTC- 1, DTC- 2, and DTC-3, are treated. The DTC-1 scheme is inspired from the Takahashi*'s strategy proposed in induction motor drives. This scheme is penalized by the high torque dips. In view of that, the DTC-2 is proposed to address the DTC-1 limitations. Simulation evaluation proves that difference between the slopes of the incoming and outgoing phases currents procreates commutation torque dips penalizing the DTC-2 scheme. This difference is more serious at high speed. To overcome the DTC-2 strategy drawbacks, a DTC-3 scheme, based on the compensation of this difference, is proposed. Taking into account the increase of this difference with the increase of the rotation speed, the vector selection table design of this strategy is synthesized to minimize the torque ripple as far as possible.*

1 Introduction

The use of three-level NPC (B12) inverter in high-power applications increases due to their performances [1]- [2]. Indeed, compared to the two-level inverter, the B12 inverter offers the low output current distortion and operates with reduced dv/dt stress [3]-[4]-[5]-[6].

Due to its numerous advantages as its simple structure, low noise, good stability and high efficiency [7]-[8]-[9], BLDC motor is extensively used in numerous industrial applications. Conversely, the torque ripple is a critical issue in BLDC motor drives [10]-[11]-[12]. Accordingly, minimizing the ripple produced in the torque is necessary to enhance the performances of BLDC motor drive system [13]-[14].

The present paper proposes two DTC strategies dedicated to the BLDC motor fed by three-level (B12) inverter. It has been found that commutation torque ripple is a main problem penalizing the first proposed strategy. This commutation phenomenon affects the industrial applications needing high performance [15]-[16]. It leads to a deteriorating efficiency of the BLCD motor. Hence, the second proposed strategy aims to address these drawbacks.

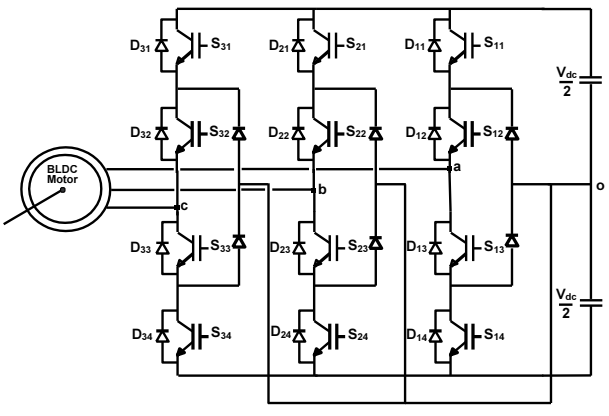


Figure 1: B12 inverter with BLDC motor connected.

*Rabiaa Mars, Sfax Engineering National School, marsrabiaa@yahoo.com

https://dx.doi.org/10.25046/aj030549

2 Analysis of the BLDC motor operation

The B12 inverter is connected with BLDC motor as illustrated in Figure.1

The rectangularly shaped currents(i_a, i_b, i_c) fed BLDC motor are synchronized with the trapezoidal back-EMFs(e_a, e_b, e_c) to accomplish a constant torque with reduced dips as shown in Figure.2.

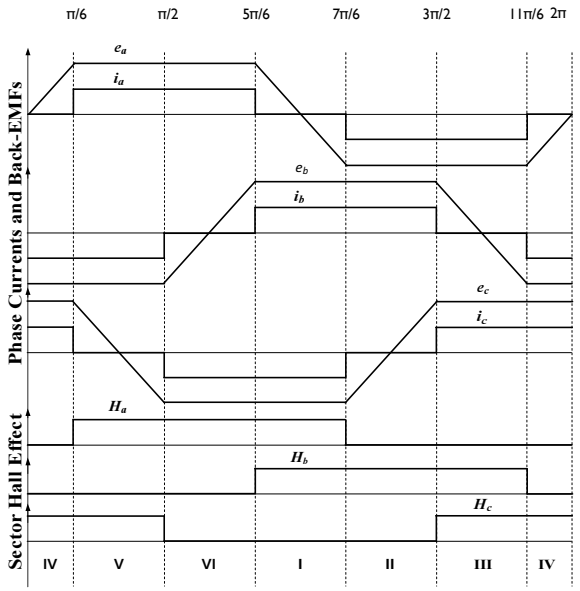


Figure 2: Back-EMFs, phase currents and the corresponding Hall-effect signals of BLDC motor .

The BLDC motor stator voltages (v_a, v_b, v_c) are given by:

$$\begin{cases} v_a = & Ri_a + (L-M)\frac{di_a}{dt} + e_a; \\ v_b = & Ri_b + (L-M)\frac{di_b}{dt} + e_b; \\ v_c = & Ri_c + (L-M)\frac{di_c}{dt} + e_c. \end{cases} \quad (1)$$

Where:

- Resistance: R (Ω),
- Self inductance: L(H),
- Mutual inductance: M(H).

The produced torque T_{em} is given as:

$$T_{em} = \frac{e_a i_a + e_b i_b + e_c i_c}{\Omega} \quad (2)$$

Where Ω represents the angular speed.

Under two-phase conduction mode, the B12 inverter engenders six Half voltage vectors (V_1 to V_6) and six Intermediate Voltage Vectors (V_7 to V_{12}).

Under three-phase conduction mode, the B12 inverter generates three null vectors (U_{01} to U_{03}), six Intermediate voltage vectors (U_1 to U_6) and six Full Voltage Vectors (U_7 to U_{12}).

Tables.1 and 2 provide the switch state combinations.

Table 1: Active voltage vectors under the two-phase conduction mode.

(S_a S_b S_c)	V_α	V_β	Vectors
Half Voltage Vectors			
(1100 0000 0000)	$\frac{1}{\sqrt{6}}V_{dc}$	0	V_1
(0000 0000 0011)	$\frac{1}{\sqrt{24}}V_{dc}$	$\frac{1}{\sqrt{8}}U_c$	V_2
(0000 1100 0000)	$-\frac{1}{\sqrt{24}}V_{dc}$	$\frac{1}{\sqrt{8}}U_c$	V_3
(0011 0000 0000)	$-\frac{1}{\sqrt{6}}V_{dc}$	0	V_4
(0000 0000 1100)	$-\frac{1}{\sqrt{24}}V_{dc}$	$-\frac{1}{\sqrt{8}}U_c$	V_5
(0000 0011 0000)	$\frac{1}{\sqrt{24}}V_{dc}$	$-\frac{1}{\sqrt{8}}U_c$	V_6
Intermediate Voltage Vectors			
(1100 0000 0011)	$\sqrt{\frac{3}{8}}V_{dc}$	$\frac{1}{\sqrt{8}}V_{dc}$	V_7
(0000 1100 0011)	0	$\frac{1}{\sqrt{2}}V_{dc}$	V_8
(0011 1100 0000)	$-\sqrt{\frac{3}{8}}V_{dc}$	$\frac{1}{\sqrt{8}}V_{dc}$	V_9
(0011 0000 0011)	$-\sqrt{\frac{3}{8}}V_{dc}$	$-\frac{1}{\sqrt{8}}V_{dc}$	V_{10}
(0000 0011 1100)	0	$-\frac{1}{\sqrt{2}}V_{dc}$	V_{11}
(1100 0011 0000)	$\sqrt{\frac{3}{8}}V_{dc}$	$\frac{1}{\sqrt{8}}V_{dc}$	V_{12}

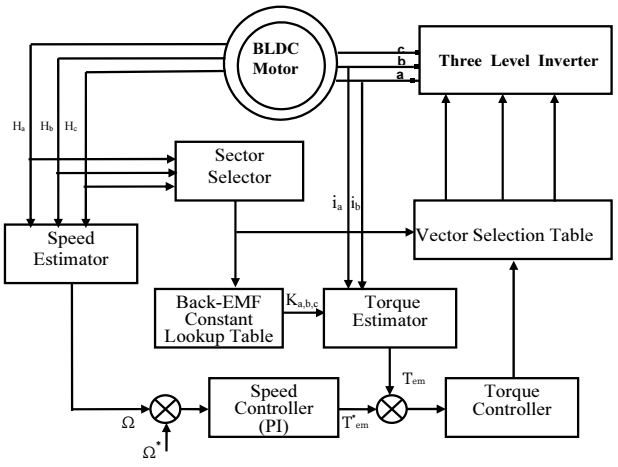


Figure 3: DTC scheme of the B12 inverter with BLDC motor connected.

3 Control strategies

3.1 DTC-1 strategy

The DTC-1 scheme is inspired from the one proposed by *Takahashi* in [17]. This strategy allows a direct control of the BLDC motor torque.

The DTC-1 scheme block diagram is illustrated in Figure.3. This diagram includes an hysteresis torque estimator which generates two outputs used to determine the vector selection table($c_\tau = +1$:torque increase, $c_\tau = -1$:torque decrease).

The switching table of this strategy is summarized in Table 3.

Table 2: Active voltage vectors under the three-phase conduction mode.

(S_a S_b S_c)	V_α	V_β	Vectors
Intermediate Voltage Vectors			
(1100 0010 0011)	$\sqrt{\frac{3}{8}}V_{dc}$	$\frac{1}{\sqrt{8}}V_{dc}$	U_1
(0100 1100 0011)	0	$\frac{1}{\sqrt{2}}V_{dc}$	U_2
(0011 1100 0010)	$-\sqrt{\frac{3}{8}}V_{dc}$	$\frac{1}{\sqrt{8}}V_{dc}$	U_3
(0011 0100 0011)	$-\sqrt{\frac{3}{8}}V_{dc}$	$-\frac{1}{\sqrt{8}}V_{dc}$	U_4
(0010 0011 1100)	0	$-\frac{1}{\sqrt{2}}V_{dc}$	U_5
(1100 0011 0100)	$\sqrt{\frac{3}{8}}V_{dc}$	$U6$	
Full Voltage Vectors			
(1100 0011 0011)	$\sqrt{\frac{2}{3}}V_{dc}$	0	U_7
(1100 1100 0011)	$\frac{1}{\sqrt{6}}V_{dc}$	$\frac{1}{\sqrt{2}}V_{dc}$	U_8
(0011 1100 0011)	$-\frac{1}{\sqrt{6}}V_{dc}$	$\frac{1}{\sqrt{2}}V_{dc}$	U_9
(0011 1100 1100)	$-\sqrt{\frac{2}{3}}V_{dc}$	0	U_{10}
(0011 0011 1100)	$-\frac{1}{\sqrt{6}}V_{dc}$	$-\frac{1}{\sqrt{2}}V_{dc}$	U_{11}
(1100 0011 1100)	$\frac{1}{\sqrt{6}}V_{dc}$	$-\frac{1}{\sqrt{2}}V_{dc}$	U_{12}

Table 3: The DTC-1 strategy vector selection table.

c_τ	+1	−1
Sector I	$\mathbf{V}_8$ (0000 1100 0011)	$\mathbf{V}_{11}$ (000000111100)
Sector II	$\mathbf{V}_9$ (001111000000)	$\mathbf{V}_{12}$ (110000110000)
Sector III	$\mathbf{V}_{10}$ (001100000011)	$\mathbf{V}_7$ (110000000011)
Sector IV	$\mathbf{V}_{11}$ (000000111100)	$\mathbf{V}_8$ (0000 1100 0011)
Sector V	$\mathbf{V}_{12}$ (110000110000)	$\mathbf{V}_9$ (001111000000)
Sector VI	$\mathbf{V}_7$ (110000000011)	$\mathbf{V}_{10}$ (001100000011)

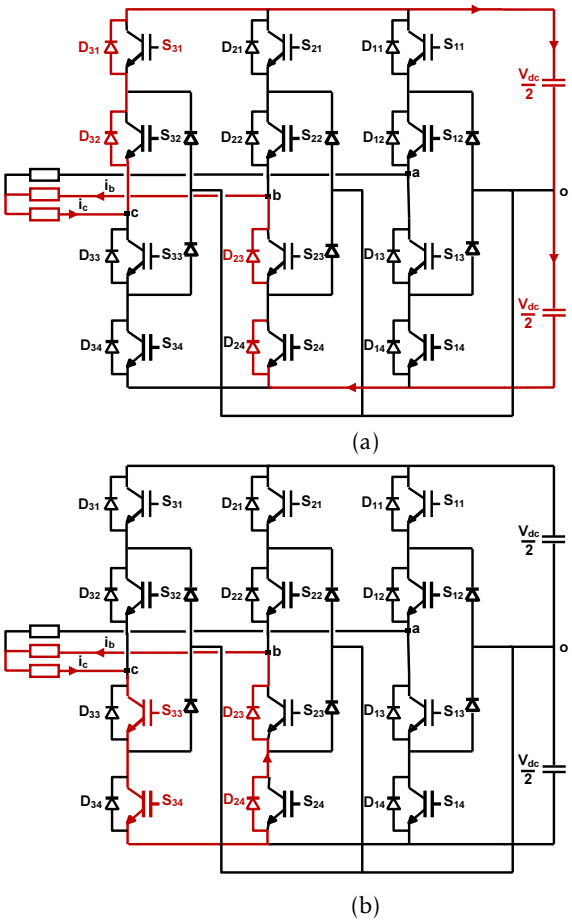


Figure 4: The current flowing during sector I in the case of a torque decrease under **Legend** (a): the DTC-1 scheme, (b): the proposed DTC-2 scheme.

3.2 DTC-2 strategy

It has been found that DTC-1 scheme is penalized by the high torque dips. In view of that, the DTC-2 is proposed to address the DTC-1 limitations. The control scheme of this strategy is similar to the one used in the DTC-1. The only difference is the control rules considered in the vector selection table.

Let's consider sector I, the vector selection table design of this scheme is synthesized as follows:

- To increase the torque, the DTC-2 scheme considers the same vectors applied in DTC-1,
- the vector V_2 achieves the torque decrease. As

illustrated in Figure. 4(b), the vector V_2 is defined by a freewheeling phase involving the two IGBTs S_{33} and S_{34} and the diodes D_{23} and D_{24}. Compared to DTC-2, the DTC-1 scheme considers the vector V_8 which characterized by a regenerative phase. During this phase, the currents i_c and i_b are forced to flow through the supply source Vdc and the diodes (D_{23}, D_{24}, D_{31} and D_{32}) as shown in Figure. 4(a) . The currents slope observed under this phase has an absolute value higher than the one obtained under the freewheeling phase.

Consequently, the resulting DTC-2 strategy vector selection table is given in Table.4.

Table 4: The DTC-2 strategy vector selection table.

c_τ	+1	−1
Sector I	$\mathbf{V}_8$(000011000011)	$\mathbf{V}_2$(000000000011)
Sector II	$\mathbf{V}_9$(001111000000)	$\mathbf{V}_3$(000011000000)
Sector III	$\mathbf{V}_{10}$(001100000011)	$\mathbf{V}_4$(001100000000)
Sector IV	$\mathbf{V}_{11}$(000000111100)	$\mathbf{V}_5$(000000001100)
Sector V	$\mathbf{V}_{12}$(110000110000)	$\mathbf{V}_6$(000000110000)
Sector VI	$\mathbf{V}_7$(110000000011)	$\mathbf{V}_1$(110000000000)

3.3 DTC-3 strategy

It has been found that DTC-2 scheme is penalized by the commutation torque ripple. As illustrated in Figure.5(a), the difference between the slopes of the incoming and outgoing phases currents generates commutation torque dips. In view of that, there is a necessity to make equal the rising time and the falling one of the phase currents as illustrated in Figure.5(b).

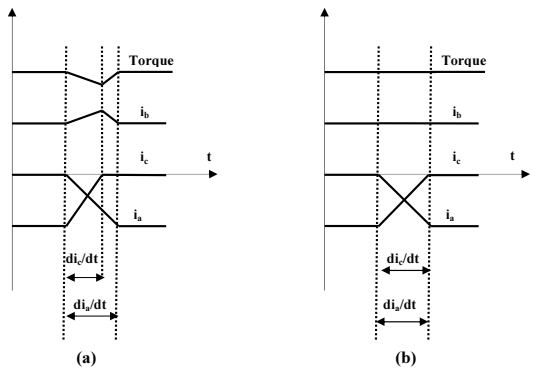


Figure 5: The behavior of the torque during commutation **Legend** (a): unequal slopes of the incoming and outgoing phases currents, (b): equal slopes of the incoming and outgoing phases currents.

The stator voltage equations of the BLDC motor are described by the equation (1). Neglecting the mutual inductance, v_a, v_b and v_c are given by the following system :

$$\begin{cases} v_a - e_a = & \mathrm{R}i_a + \mathrm{L}\frac{di_a}{dt}\ ; \\ v_b - e_b = & \mathrm{R}i_b + \mathrm{L}\frac{di_b}{dt}; \\ v_c - e_c = & \mathrm{R}i_c + \mathrm{L}\frac{di_c}{dt}. \end{cases} \tag{3}$$

The Laplace Transform method can be used to solve this linear differential equations. Therefore the relation between the phase voltage and the phase current is dealt with using this method, and then the transfer function is similarly written as:

$$H(p) = \frac{v(p) - e(p)}{i(p)} = \frac{1}{Lp - R} \tag{4}$$

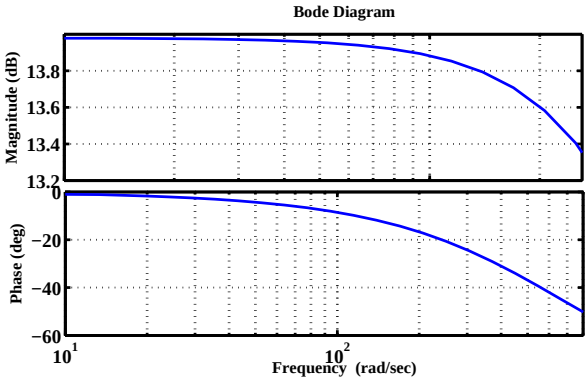


Figure 6: Approximated curve of the frequency response (bode-plot) of the BLDC motor.

Figure6 shows the approximated curve of the frequency response (bode-plot) based on equation (4). The plot displays the magnitude (in dB) and the phase (in degrees) of the system response as a function of frequency. Referring to this figure, the phase delay angle increases with the increase of the rotation speed. Moreover, this delay is more important at high speed. Consequently, the commutation torque dips is very serious at high speed.

To address the DTC-2 limitations, the third DTC scheme, noted DTC-3, is proposed . This strategy is based on the employ of a three-level torque controller. This controller generates 3 outputs as shown in Figure.7 :

- -1 corresponding to the torque decrease,
- +1 corresponding to the torque increase,
- +2 is activated when the torque drops during commutations from sector-to-sector .

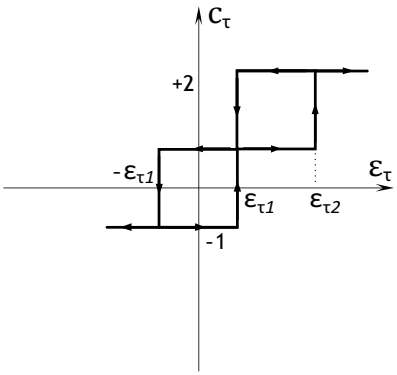


Figure 7: The three-level-torque controller characteristics.

Let's consider sector I, the vector selection table design of this scheme is synthesized as follows:

- c_τ=+1 and c_τ=-1 are achieved by the same vectors used in DTC-2 scheme.
- The suppression of the commutation torque dips (c_τ=+2) is achieved by the application of the vectors corresponding to the three-phase conduction mode. To equalize $|di_a/dt|$ and $|di_c/dt|$,at the beginning of sector II, in the case of a reference speed $\Omega^* \leq \Omega_{limit}$, the current i_c is forced to flow through the IGBT S_{33}, the clamped diode and the supply source $\frac{V_{dc}}{2}$ instead of the uncontrollable diodes (D_{31} and D_{32}) as shown in Figures. 8(b) and 8(c). For a reference speed $\Omega > \Omega_{limit}$, the current i_c is forced to flow through the IGBTs (S_{33} , S_{34}) and the supply source V_{dc} instead of the uncontrollable diodes (D_{31} and D_{32}) as shown in Figures. 8(b) and 8(d).

Consequently,the resulting vector selection is given in Table.7.

4 Evaluation of the proposed DTC schemes effectiveness and performances

Simulation work has been accomplished to highlight the proposed strategies performances. Tables.5 and 6 gives the BLDC motor parameters and ratings.

Table 5: The ratings of BLDC motor

Current = 30A	DC voltage = 24V
Rated torque = 2.5Nm	PM flux-linkage = 14mWb
Power = 600W	Rated speed = 2500rpm

Table 6: BLDC motor parameters

R = 0.2Ω	p = 3
L = 0.3mH	J = 4.1g.m^2
R = 0.2Ω	f = 0.5mN.m.s

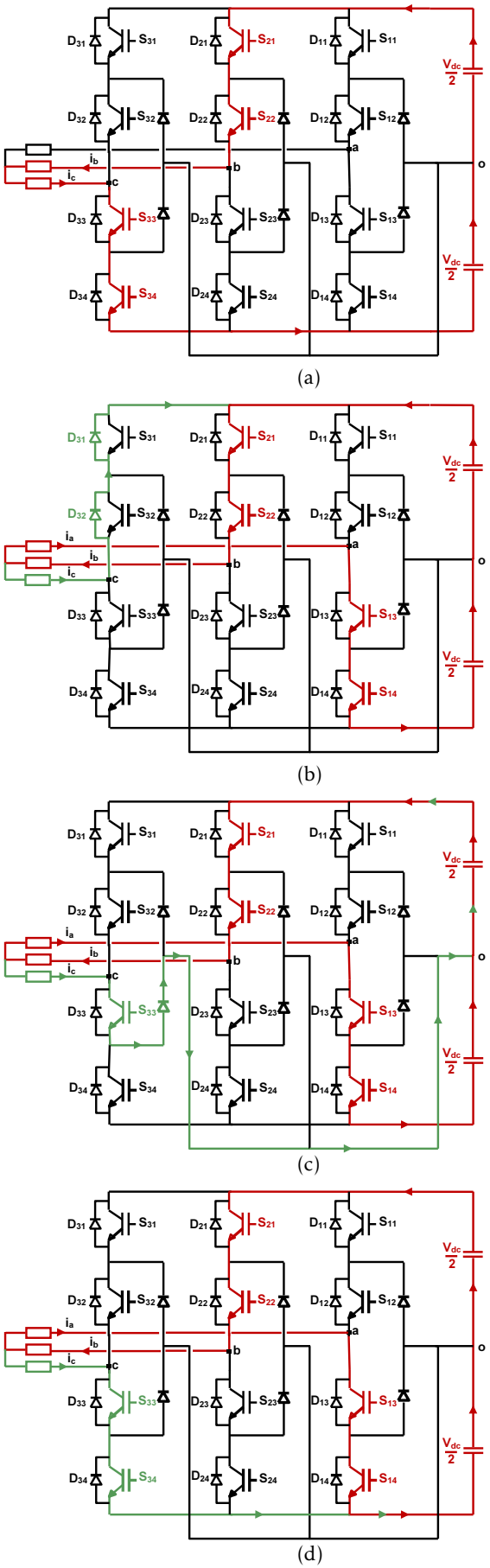


Figure 8: Current flowing during : **Legend** (a): sector I, (b): commutation under DTC-2 scheme, (c): commutation under DTC-3 scheme($\Omega^* \leq \Omega_{limit}$) and (d): commutation under DTC-3 scheme ($\Omega^* > \Omega_{limit}$).

Table 7: The vector selection table of the DTC-3 strategy.

c_τ	+1	−1	c_τ	+2	
				$\Omega^* \leq \Omega_{limit}$	$\Omega^* > \Omega_{limit}$
Sector I	$\mathbf{V}_8$(000011000011)	$\mathbf{V}_2$(000000000011)	Sector VI → Sector I	$\mathbf{U}_2$(010011000011)	$\mathbf{U}_8$(110011000011)
Sector II	$\mathbf{V}_9$(001111000000)	$\mathbf{V}_3$(000011000000)	Sector I → Sector II	$\mathbf{U}_3$(001111000010)	$\mathbf{U}_9$(001111000011)
Sector III	$\mathbf{V}_{10}$(001100000011)	$\mathbf{V}_4$(001100000000)	Sector II → Sector III	$\mathbf{U}_4$(001101000011)	$\mathbf{U}_{10}$(001111001100)
Sector IV	$\mathbf{V}_{11}$(000000111100)	$\mathbf{V}_5$(000000001100)	Sector III → Sector IV	$\mathbf{U}_5$(001000111100)	$\mathbf{U}_{11}$(001100111100)
Sector V	$\mathbf{V}_{12}$(110000110000)	$\mathbf{V}_6$(000000110000)	Sector IV → Sector V	$\mathbf{U}_6$(110000110100)	$\mathbf{U}_{12}$(110000111100)
Sector VI	$\mathbf{V}_7$(110000000011)	$\mathbf{V}_1$(110000000000)	Sector V → Sector VI	$\mathbf{U}_1$(110000100011)	$\mathbf{U}_7$(110000110011)

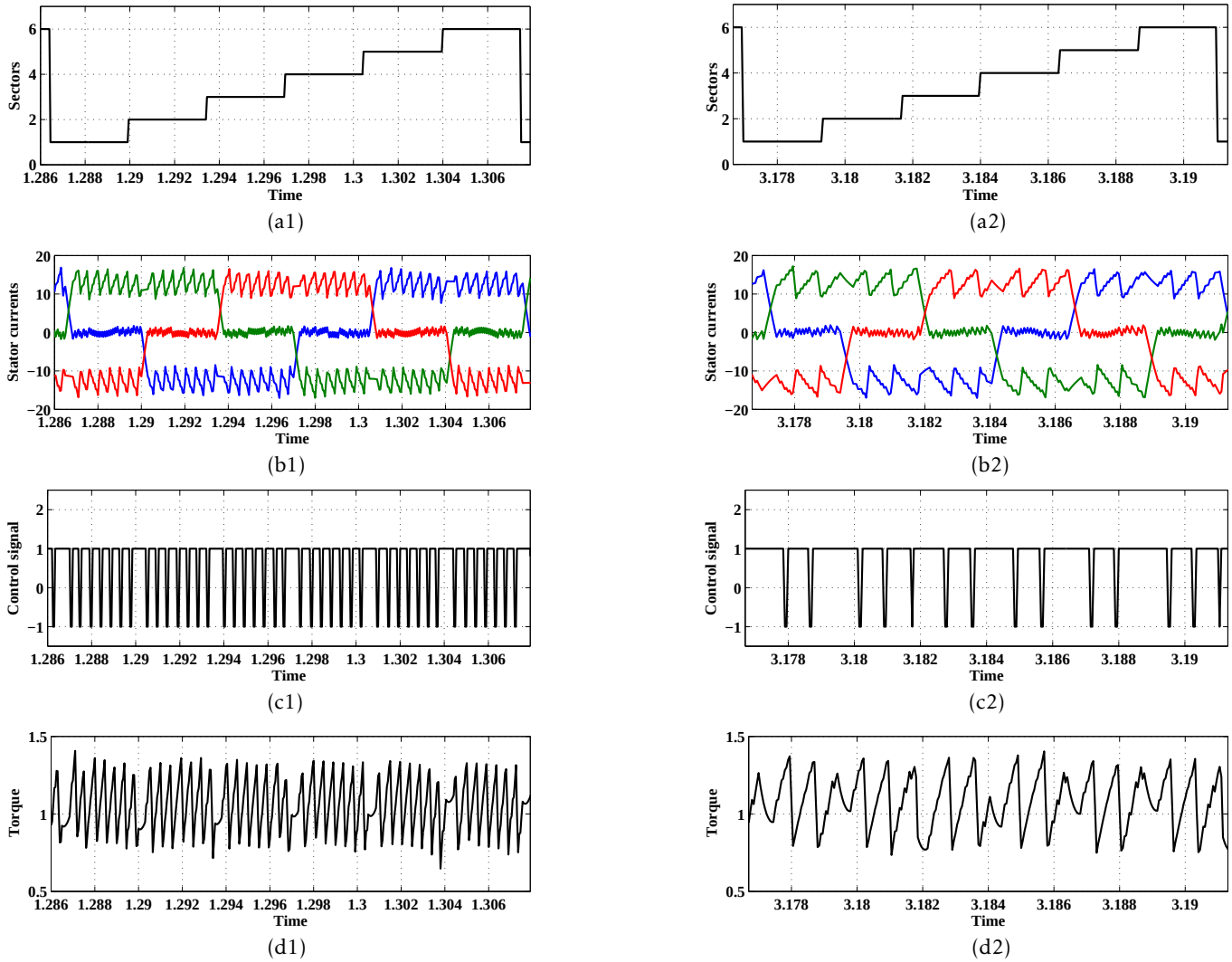


Figure 9: Simulated steady-state variables provided by the classic DTC-1 scheme at Ω = +100rad/s (subscript "1") and Ω = +150rad/s (subscript "2"). **Legend** (a): sector succession, (b): stator currents, (c): control signal c_τ and (d): electromagnetic torque.

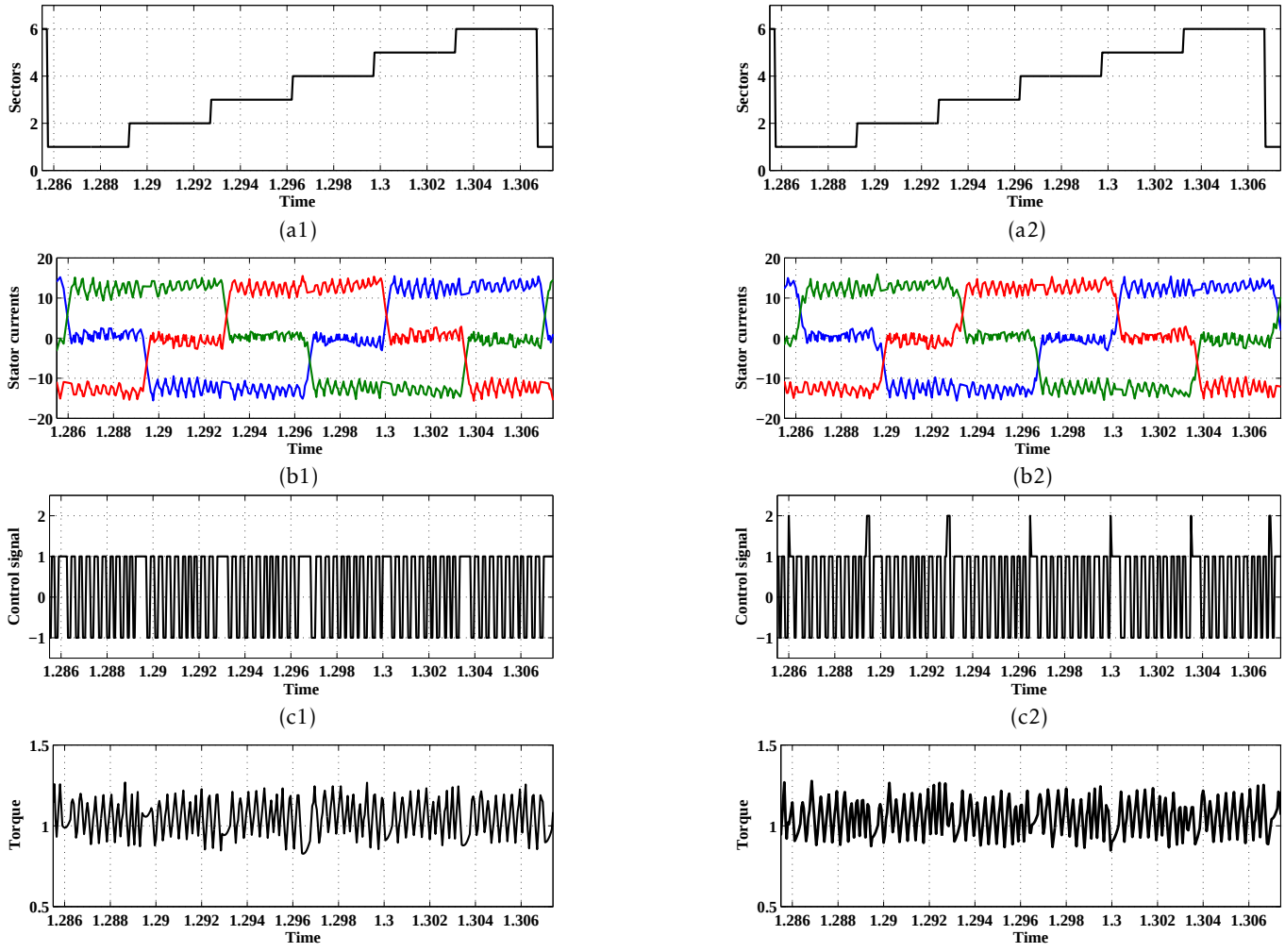


Figure 10: Simulated steady-state variables offered by the DTC-2(subscript "1") and the proposed DTC-3 (subscript "2") schemes at Ω = +100rad/s. **Legend** (a): sector succession, (b): stator currents, (c): control signal c_τ and (d): electromagnetic torque.

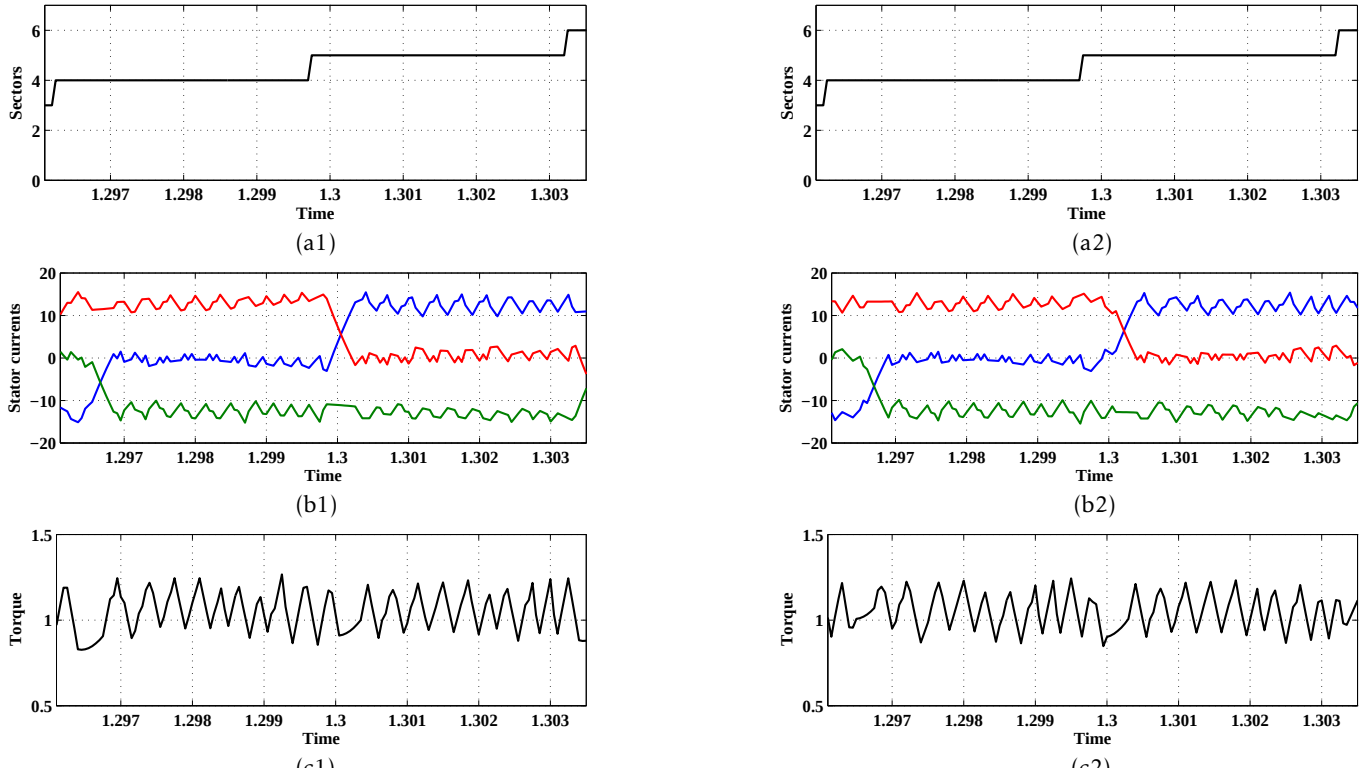


Figure 11: Simulated steady-state variables offered by the DTC-2 scheme (subscript "1")and DTC-3 (subscript "2") strategies for a speed Ω =+ 100rad/s. **Legend** (a): commutation from sector IV to sector V, (b): the stator currents zoom , and (c): the electromagnetic torque zoom.

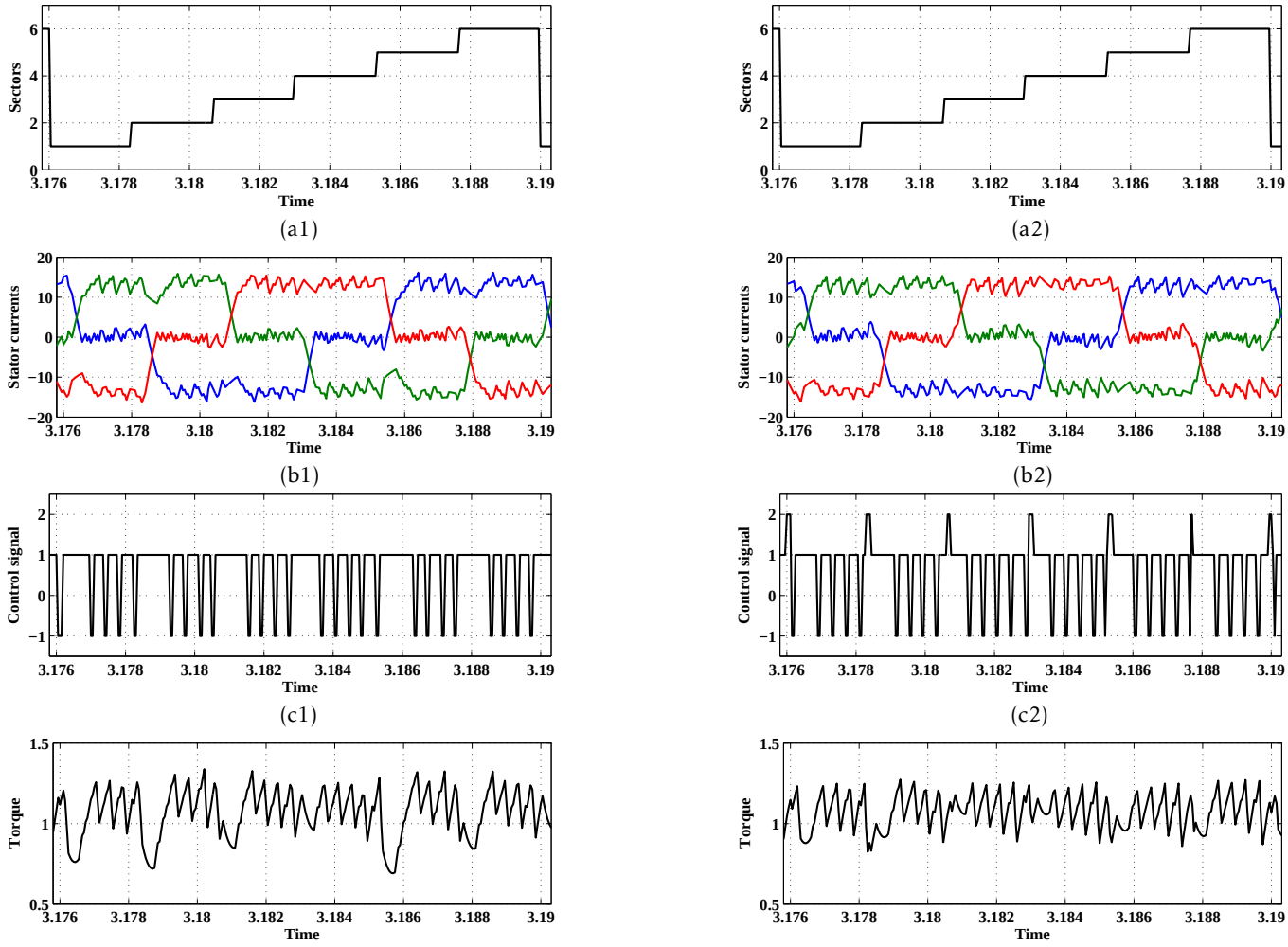


Figure 12: Simulated steady-state variables yielded by the DTC-2 (subscript "1") and the proposed DTC-3 strategies at $\Omega = +150$rad/s. **Legend** (a): sector succession, (b): stator currents, (c): control signal c_τ and (d): electromagnetic torque.

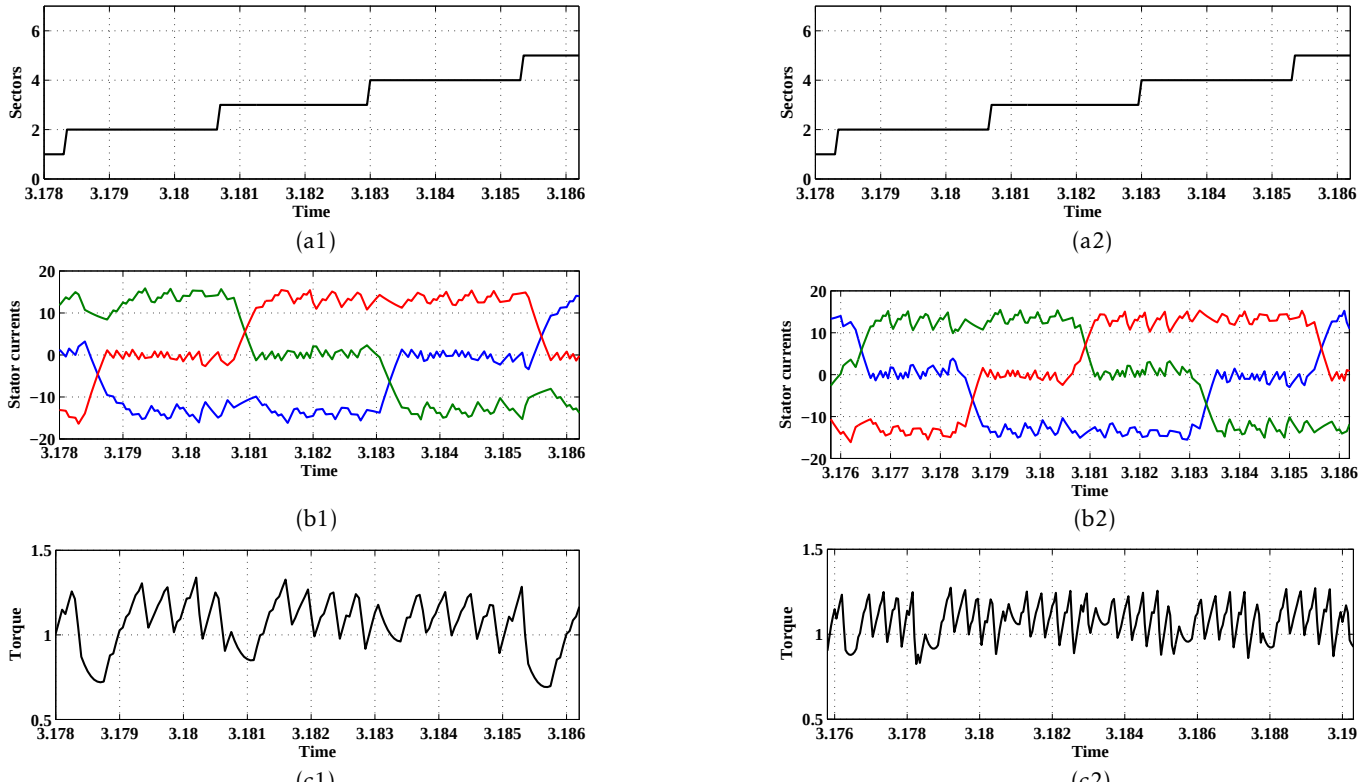


Figure 13: Simulated steady-state variables yielded by the DTC-2 strategy (subscript 1) and the proposed DTC-3 one for a speed $\Omega = +150$rad/s . **Legend** (a): commutations from sector II to sector IV, (b): the stator currents zoom, and (c): the electromagnetic torque zoom .

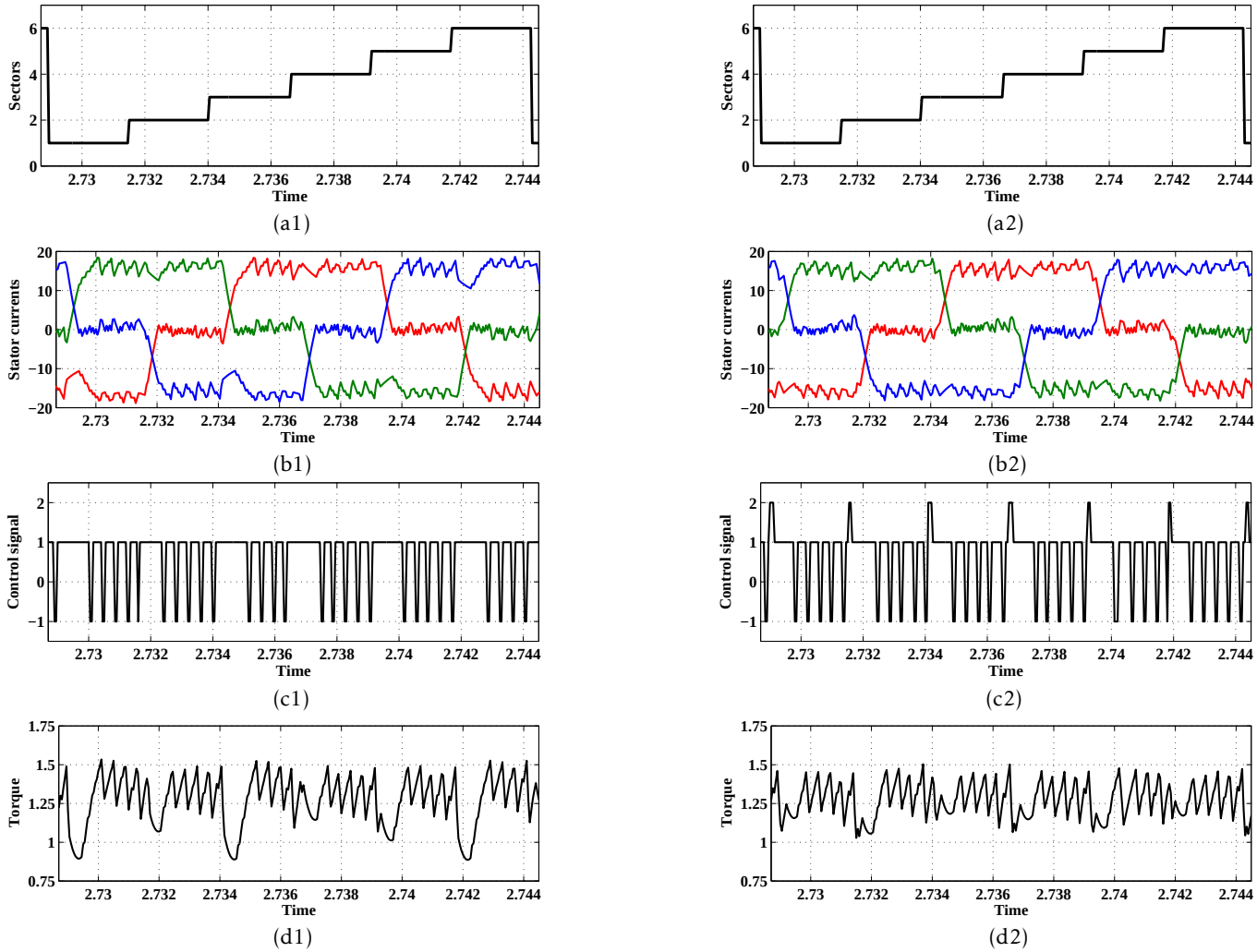


Figure 14: Simulated transient-state variables offered by the DTC-2(subscript "1") and the proposed DTC-3 schemes. **Legend** (a): sector succession, (b): stator currents, (c): control signal c_τ and (d): electromagnetic torque.

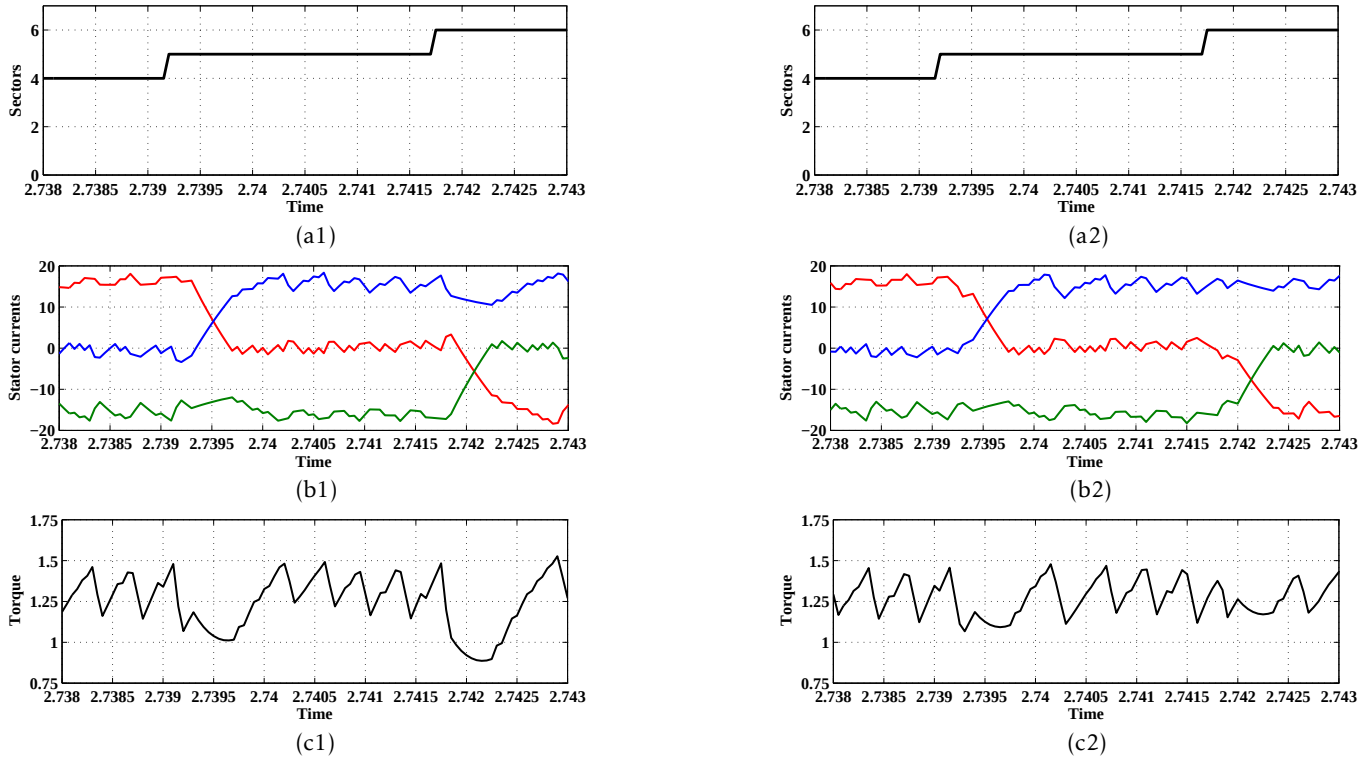


Figure 15: Simulated transient-state variables offered by the DTC-2 strategy (subscript 1) and the proposed DTC-3 one.**Legend** (a): commutations from sector V to sector VI, (b): the stator currents zoom , and (c): the electromagnetic torque zoom .

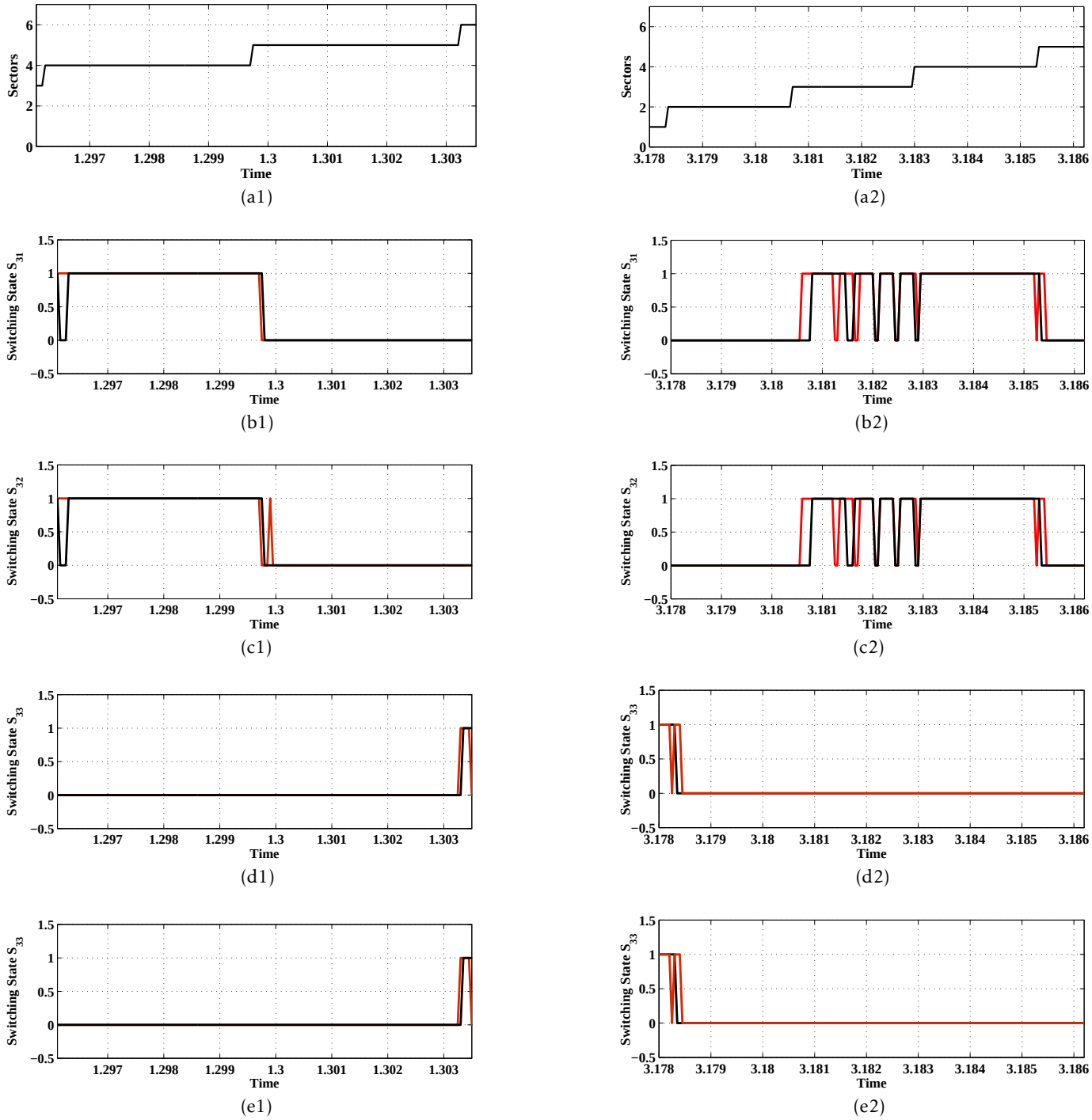


Figure 16: Simulated steady-state variables under the DTC-2 scheme (color black) and the proposed DTC-3 one (color red) for two reference speeds Ω =+ 100rad/s (subscript 1) and Ω =+ 150rad/s (subscript 2). **Legend** (a): commutations from sector II to sector IV, (b): S_{31}, (c): S_{32}, (d): S_{33}, and (e): S_{34} .

Moreover, simulation conditions characterizing the simulation study are listed below:

- a sampling period T_s=55μs,
- a proportional gain $K_p = 1$,
- an integral gain $K_i = 0.5$,
- a load torque = 1Nm,
- a limit speed $\Omega_{limit} = 110 rad/s$
- the hysteresis bounds are: $\varepsilon_{\tau 1} = \pm 0.0012$Nm and $\varepsilon_{\tau 2} = \pm 0.0036$.

To analyze the effectiveness of the proposed DTC-2 and DTC-3 strategies, this section deals with a simulation-based comparison between these strategies. By using MATLAB/SIMULINK, the proposed strategies are implemented. The obtained simulations results are illustrated in Figures. 10, 11, 12, 13, 14, 15.

4.1 Performance of the proposed strategies during sequence operation:

Figures.9, 10 and 12 show the waveforms of the current, the torque controller output c_τ and the torque waveforms at steady-state operation.

By comparing Figures.9(d), 10(d) and 12(d), one clearly notice that the DTC-1 scheme leads to an important torque dips thanks to the application of the Intermediate active voltage vectors. Both DTC-3 and DTC-2 strategies lead to a low torque dips during sequence operation thanks to the application of the Half active voltage vectors.

During sequence operation, both DTC-2 and DTC-3 schemes present almost the same torque waveforms. This is expected thanks to the application of the the same vectors. The difference concerns the torque decrease during commutations from sector to sector .

4.2 Performance of the proposed strategies during commutations from sector-to-sector :

To evaluate the proposed DTC schemes effectiveness during commutations from sector-to-sector , Figures. 11 and 13 give a zoom of the electromagnetic torque and the stator currents at steady-state operation. Figure.16 illustrates the steady-state waveforms of the switching states S_{31}, S_{32}, S_{33} and S_{34} with the sector succession.

As shown in Figures. 11(b1) and 13(b1), during the commutation from sector IV to sector V, the unbalancing between the current rising time and the current falling time affected the electromagnetic torque waveform. As a result of this, the DTC-2 strategy is penalized by a high torque dips. The comparison of Figures.10(c1) and 12(c1) confirms that commutation torque dips is more serious and more strong at high speed.

Compared to DTC-2, this commutation phenomenon is resolved for the DTC-3 strategy. Therefore, during the commutation from sector IV to sector V and in the case of the reference speed $\Omega^* = 150 rad/s$ ($\Omega^* > \Omega_{limit}$), the current i_c is forced to flow through the IGBTS (S_{31}, S_{32}) and the supply source V_{dc} instead of the freewheeling diodes (D_{33} and D_{34}) as shown in Figures.16(b1), (b2), (b3) and (b4). However, from the analysis of Figures.16(a1), (a2), (a3) and (a4) in the case of a reference speed $\Omega^* = 100 rad/s$ ($\Omega^* \leq \Omega_{limit}$), it can be clearly noticed that the current i_c is forced to flow through the IGBT S_{32}, the clamped diode and the supply source $\frac{V_{dc}}{2}$ instead of the freewheeling diodes (D_{33} and D_{34}). Consequently, $|di_a/dt|$ and $|di_c/dt|$ are equal for both reference speeds which results a high dynamic of the torque without commutation ripple. These benefits are confirmed in Figures. 11(a2), 11(a3), 13(b2) and 13(b3).

The performance of DTC strategies under a transient-state is presented in Figures.14 and 15. Accordingly, compared to DTC-2 and DTC-1 schemes, one can evidently affirm that the developed DTC-3 scheme leads to the lower commutation torque dips for both transient and steady-state operations.

5 Conclusion

BLDC motor is integrated in various applications due to its different advantages. However, the BLDC motor drive has the drawback of high torque ripple which affects the performance of this system. Thus, torque ripple minimization is a crucial necessity. This paper has been focused on the development of two DTC schemes dedicated to BLDC motor drives. Simulation work has been accomplished to highlight the developed schemes effectiveness and performances. It has been found that the developed DTC-3 scheme has guaranteed the less distortion in the line currents and less commutation torque dips. This developed strategy has been based on the compensation of the difference between the slopes of the incoming and outgoing phases currents.

References

[1] J. Rodriguez, J.-S. Lai, and F. Z. "Peng, Multilevel inverters: A survey of topologies controls and applications", *IEEE Trans. Ind. Electron.*, vol. 49, no. 4, pp. 724-738, Aug. 2002.

[2] Y. Liu, A. Q. Huang, W. Song, S. Bhattacharya, and G. Tan, "Smallsignal model-based control strategy for balancing individual dc capacitor voltages in cascade multilevel inverter-based statcom, Industrial Electronics",*IEEE Transactions on*, vol. 56, no. 6, pp. 22592269, 2009

[3] H. Vahedi,K. Al-Haddad, P.-A. Labbe, and S. Rahmani, "Cascaded multilevel inverter with multicarrier pwm technique and voltage balancing feature", *in Industrial Electronics (ISIE)*, 2014 IEEE 23rd International Symposium on. IEEE, 2014, pp. 21552160

[4] J. P. Lyons, V. Vlatkovic, P. M. Espelange, A. A. M. Esser, and F. F. Want, "Five level high power motor drive converter and control system", *U.S. Patent 6 058 031*, May 2, 2000.

[5] S. Bernet, T. Bruckner, and P. Stiemer, Three-point converter and method for its operation, U.S. Patent 6 219 265, Apr. 17, 2001.

[6] M. A. Doss, E. Premkumar, G. R. Kumar, and J. Hussain, "Harmonics and torque ripple reduction of Brushless DC Motor (BLDCM) using cascaded H-bridge multilevel inverter" , *in Proc. IEEE ICPEC*, Feb. 2013, pp. 296299.

[7] Chan, C. C., Jiang, J. Z., Xia, W., Chan, K. T. "Novel wide range speed control of permanent magnet brushless motor drives". *IEEE transactions on power electronics*, 10(5), 539-546,1995

[8] T. Shi, Y. Guo, P. Song, and C. Xia, "A new approach of minimizing commutation torque ripple for brushless DC motor based on DC-DC converter", *IEEE Trans. Ind. Electron.*, vol. 57, no. 10, pp. 3483-3490, Oct. 2010.

[9] K. Ohishi, M. Nakao, K. Ohnishi, and K. Miyachi, "Microprocessor-controlled DC motor for load-insensitive position servo system", *IEEE Trans. Ind. Electron.*, vol. 34, no. 1, pp. 44-49, 1987.

[10] X. Huang, A. Goodman, C. Gerada, Y. Fang, and Q. Lu, "A single sided matrix converter drive for a brushless dc motor in aerospace applications", *IEEE Trans. Ind. Electron.*, vol. 59, no. 9, pp. 3542-3552, Sep. 2012.

[11] T. M. Jahns and W. I. Soong, "Pulsating torque minimization techniques for permanent magnet AC motor drives-a review", *IEEE Trans. Ind. Electron.*, vol. 43, no. 2, pp. 321-330, Apr. 1996.

[12] X. Li, C. Xia, Y. Cao, W. Chen, and T. Shi, "Commutation torque ripple reduction strategy of Z-source inverter fed brushless DC motor", *IEEE Trans. Power Electron.*, vol. 31, no. 11, pp. 7677-7690, Nov. 2016.

[13] C. Xia, Y. Xiao, W. Chen, and T. Shi, "Torque ripple reduction in brushless DC drives based on reference current optimization using integral variable structure control", *IEEE Trans. Ind. Electron.*, vol. 61, no. 2, pp. 738-752, Feb. 2014.

[14] G. Buja, M. Bertoluzzo, and R. K. Keshri, "Torque ripple-free operation of PM BLDC drives with petal-wave current supply", *IEEE Trans. Ind. Electron.*, vol. 62, no. 7, pp. 4034-4043, Jul. 2015.

[15] J. Fang, H. Li, and B. Han. "Torque ripple reduction in BLDC torque motor with non-ideal back EMF", *IEEE Trans. Power Electron.*, vol. 27, no. 11, pp. 4630-4637, Nov. 2012.

[16] W. Jiang, Y. Liao, J. Wang, P. Wang, and Y. Xie, "Improved control of BLDCM considering commutation torque ripple and commutation time in full speed range", *IEEE Trans. Power Electron.*, vol. 33, no. 5,pp. 4249-4260, June 2018.

[17] I. Takahashi and T. Noguchi, "A new quick-response and high efficiency control strategy of an induction motor", *IEEE Trans. Ind. Appl.*, vol. 22, no. 5, pp. 820-827, 1986.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 432-437 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

An Integrated & Secure System for Wearable Devices

Callum Owen-Bridge[1], Stewart Blakeway[1,2], Emanuele Lindo Secco[1,*]

[1]*Robotic Laboratory, Department of Mathematics and Computer Science, Liverpool Hope University*

[2]*Faculty of Arts, Science and Technology, Wrexham Glyndŵr University*

ARTICLE INFO

Article history:
Received: 13 September, 2018
Accepted: 27 November, 2018
Online: 19 December, 2018

Keywords:
E-Health
Wearable sensors
Human Health Monitoring

ABSTRACT

Health services are under increasing pressure to reduce large waiting times for appointments. The delayed diagnosis of a human illnesses could have profound consequences for the patient. The large waiting times may be attributed to lack of data, the accuracy of the data or the timeliness of the availability of the data. In addition to this people often priorities their busy lives over their health. Other factors that affect delayed diagnosis of patients could be attributed to ill-health and the difficulty in the patient visiting the health-care practitioner, this is more evident in elderly patients that may have suffered a deterioration of their health. To overcome these obstacles eHealth systems could be put in place to facilitate the transmission of real-time and accurate data regarding a patient's health directly to secure database that can be accessed by authorized practitioners. This paper outlines the use of eHealth (electronic health) to improve health care services which makes use of a wearable set of secure sensors and a secure database that is accessible using a web application. The proposed wearable system records and stores encrypted data which is related to the body's physical parameters. Then, the encrypted data are transmitted using wireless technologies from the wearable device to a secure relational multi-user database using a protected protocol. The data can be accessed by the patient and professionals, such as general practitioners, using a web interface once authentication has taken place. This secure wearable architecture alleviates the problems of the lack of data, the inaccuracy of data and the timeliness of data by recording vital body parameters throughout the day and by sending real-time live data to a system that can be immediately accessed by the practitioner. The system also allows for the automatic analysis of data and presentation of the data using graphs which could aid the practitioner in determining patterns in the patients' health statistics.

1. Introduction

Health care services are struggling to diagnose people accurately and quickly, due to lack of data. Humans are becoming less proactive and unaware of their own health. Many people struggle to allocate time to see a health care professional due to their lifestyle and/or physical health.

This research presents a secure medical wearable device for patients with health issues, to help support in the monitoring of physical parameters in the daily life and in the Ambient Assisted Living (AAL) context. This paper focuses on the creation and implementation of a hardware device with sensors that captures patients health data and the development of secure software for data storage which can be accessed remotely by multiple authorised users.

The research undertaken details how the hardware was implemented, the software used to control the device, how data is secured before being transmitted over unsecured wireless media to a secure multiple-access relational database. This papers also discusses eHealth, health of the population, security, privacy and legal issues surrounding medical wearable devices. The result of this research shows that there are high expectations and standards for eHealth strategies. Providing that security and privacy mechanisms are put in place, there seems high motivation in the adoption of such systems.

*Emanuele Lindo Secco, Robotic Laboratory, Department of Mathematics & Computer Science, Liverpool Hope University, L16 9JD, UK, +44 (0) 151 291 3641, seccoe@hope.ac.uk

The construction of the secure medical wearable device and development of the secure database with the front-end website application is outlined in the materials and methods section. The results of this research show that secure wearable devices for this purpose can be successful and the results of the application of these devices is discussed in the discussion and conclusion sections.

1.1. Population Health

It is reported that across the globe the causes of death in 2015, are related to cardiovascular and respiratory systems, these figures have risen since the year 2000 [1], for example, ischaemia heart disease had risen by 2 million in 2015 [1]. In 2015, diabetes mellitus caused around 1.7 million deaths and became one of many new causes of death and diseases related to the respiratory system caused around 1.8 million deaths [1].

A medical device which could record body parameters especially those parameters involving the cardiovascular system and respiratory system would be beneficial in monitoring the health of the population and improving healthcare. The device could provide crucial data to help reduce illness and diseases. The system could also be used to identify detestation of health so that preventative measure can be put in place.

1.2. eHealth

The *eHealth* field involves public health and medical informatics; it describes the enhancement of health services through use of technology [2].

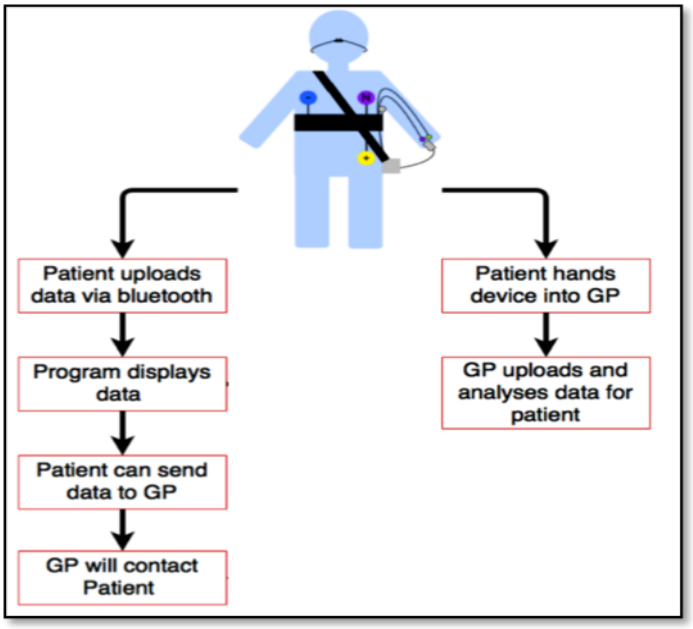


Figure 1: Functional overview of the Health Monitoring System

Research has shown that *eHealth* has provided improved support for patients whilst reducing the overall financial cost and improving the efficiency of healthcare. Portable health monitors have shown to increase self-care, improve quality of life and assist in providing better medical care [3]. An eHealth system can provide multiple benefits to the patient. The reasons for this are:

- Increasing the efficiency of health care by improving communication between the healthcare professionals and the patients. Prevention of data redundancy, increased availability of data and improved accuracy of the data play an important role in the correct diagnosis of patients.
- eHealth strategies should be evaluated scientifically for its effectiveness and efficiency [4] which allow for improved services to the patient.
- eHealth improves time efficiency of health care, services such as telehealth save time because appointments can be made online or consultations could take place using video streaming services over the internet.
- eHealth also allows people to be more aware of their health and allows individuals to easily share data with their health care provider. Which provides for more effective and efficient health care because patients can be diagnosed more quickly and accurately.
- An eHealth System can also reduce administration as the health data can be digitally securely before being shared. This reduces hard copies (paper) which could be stored insecurely or even lost.
- Moreover, electronic systems could have validation built in which may reduce some human administrative errors resulting in more accurate data.
- Patients that are concerned with the security or the privacy of their data can be assured that the data is governed by The EU General Data Protection Regulation.
- Software for such systems is becoming easier to use. The technology should be simple to use and user-friendly [5] for accessibility to the patient.

eHealth is a successful mechanism in solving many health care issues. These innovative solutions could also be invaluable for developing countries by reducing the administrative burden involved in providing health care.

The eHealth platform has shown to be a widely used health care scheme by many countries and the use of eHealth is increasing as more countries are adopting the scheme. Since 1990 there has been a steady adoption of eHealth [6], more recently application of eHealth has seen a dramatic increase. Health care professionals usually assess a patient's condition at regular intervals, which results in discreet data. eHealth devices provide continuous data which gives a more complete medical history, this could help health care professionals make better informed decisions regarding treatments for patients. This is crucial as the patient could form new symptoms between the intervals the health care professional assesses the patient [7].

1.3. Security, Privacy and Legal considerations surrounding Medical Wearable Devices

There are security, privacy and legal issues surrounding eHealth wearable devices which must be addressed for the service to be viable. It is essential that the devices are kept secure because the recorded data hold personal information of patients. If security were compromised there could be catastrophic consequences, for example, the patient being incorrectly diagnosed or receiving incorrect treatment, which may result in the death of the patient.

Of course, this is true of all systems that store patient data and procedures must be in place to secure these systems.

The proposed medical wearable device in this research does transmit data over an insecure wireless medium and this connection link could be eavesdropped [8], but this is unlikely because of the proximity of the transmitter and receiver and the likely distance the signal will travel from the low powered transmitter (theoretically the signal should be confined to about 10m). However, even with such a low risk of interception of data, steps are in place to ensure that the data is encrypted prior to transmission. The data are also stored in a secure database in encrypted form and can only be accessed by authorised personal using the authentication scheme.

Listed below are some security and privacy concerns with medical wearable devices:

- The device's system is vulnerable to eavesdropping, denial of service, unauthorized access to the device and data mining [9].
- In line with GDRP (General Data Protection Regulation), only required data should be stored. In addition to this the data stored should not allow an adversary to determine the identity of the patient.

The use of authentication mechanisms prevents unauthorized access to the data, examples of these are usernames, digital signatures and challenge-response protocols [10].

unique identification numbers which hides identifiable patient data and prevents patient identification. There is an authentication scheme and an access control list policy which prevents anyone without the correct authentication credentials from accessing the database.

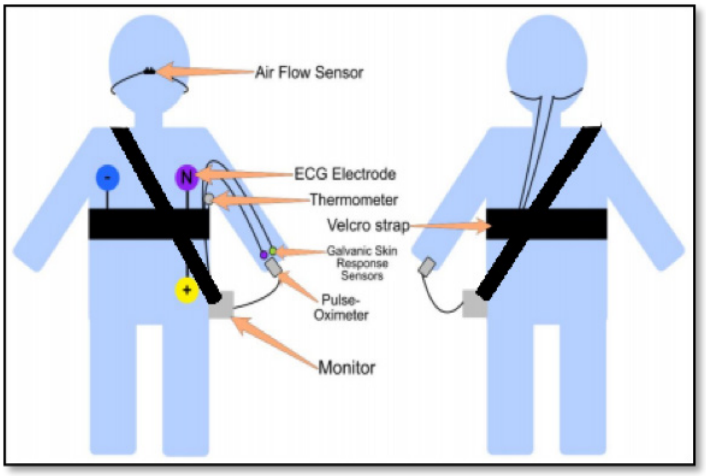


Figure 2 – Sensor configuration of the Health Monitoring System

2. Materials & Methods

The overall system process shown in Figure 1 and the creation of the secure medical wearable device shown in Figure 2 and web application shown in Figure 4, will be outlined within this chapter. The goal was to create a secure medical wearable device for use on humans, to allow for independent health monitoring, health awareness, and improved health care, through use of eHealth.

2.1. Secure Medical Wearable Device

The wearable device was made up of an Arduino UNO Board, eHealth shield [11], data-logger shield and Bluetooth shield. The eHealth shield was used as it allows the Arduino to record and measure data of the human body parameters.

The sensors used with the eHealth shield are reported in Figure 2 and 3, which are:

- Electro cardiogram (ECG) and Electromyograph (EMG): A three electrode sensor. The electrodes can be used to either record data for an ECG or EMG. The ECG records data for heart health and the EMG records data for muscle health.
- Pulse-Oximeter: A non-invasive small battery-operated device which attaches to the index finger. Records blood oxygen levels and heart rate.
- Glucometer: Records blood glucose in Millimoles per litre.
- Body Thermometer: placed at the axillary point of the body. Records body temperature.
- Air Flow: Measures the breathing rate and is placed on the upper lip.
- Galvanic Skin Response: measures the electrical conductance of the skin. This consists of two sensors placed on two separate fingers.

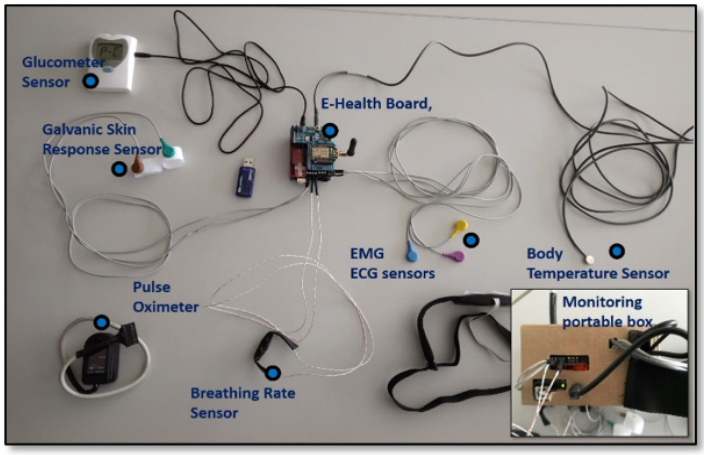


Figure 3 – The sensors and monitoring box set-up of the Health Monitoring System

The data logger shield allows the Arduino to store all the measured data onto an SD Card, with the correct date due to an

onboard real-time clock. This is programmed to store data every minute. Each log contains the date, time and the sensor readings. An authentication method was added to the SD card. To ensure the SD card was correct for the patient, so the data stored on the SD card had to match the data on the device before sensor recording would take place. The Bluetooth shield allows the Arduino to transfer data via a different medium. This shield uses a Bluetooth communication protocol (IEEE 802.15.1, Bluetooth 2.1 +EDR Class 2) and Asynchronous Encryption Standard with data transfers.

The Arduino board was used as it allows for these shields to be used and stacked together.

The Arduino Boards and shields was then encased within a box for protection of the user and the device. The box allowed access to ports and cables. A Velcro strap was attached to the box, so it can be worn over the shoulder and keep the cables tidy and managed. The strap also went across the chest supporting the ECG and thermometer cables. The device was powered by a 1200 mAh battery, with an output of 0.8 A and 5V. The power bank included LEDs to indicate charging and power.

2.2. Communication with the Device

An application was developed using Processing IDE, to communicate with the health monitor via Bluetooth or USB cable. This application was programmed to search for the Arduino board, once found it would display a window asking to name a file for which the data will be stored in. Once entered, three coloured steps were added to retrieve data, close the file of the stored data and reset the device.

2.3. Web application mask

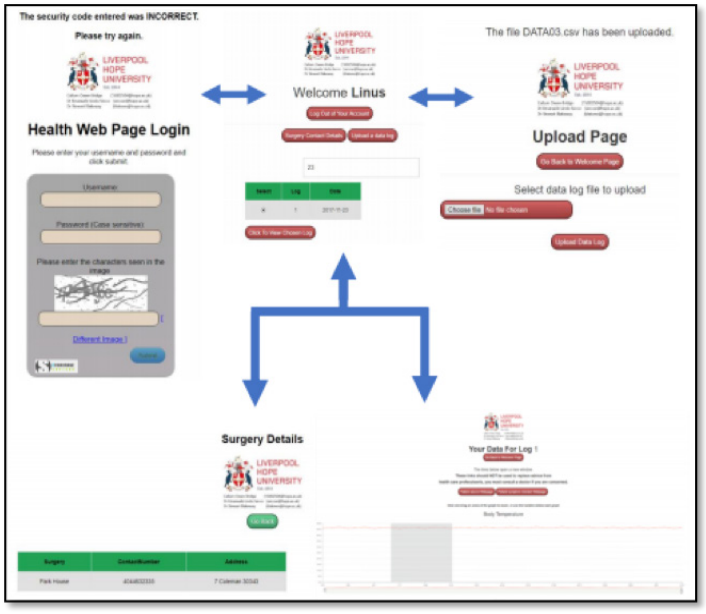


Figure 4 - The Web Application Masks for the Patient

A *database* was also designed to store the patient information, the health care professional information and the data output of the health monitoring system. The database was stored on a WampServer version 3.1.0 [12]. This server allowed storage of a MySQL database, web application and the communication between them. This database held login information using SHA2 at 512 bits to hash passwords for the web application.

A *website* was also designed. A login page was added, containing field box for username, password and secure image text. Depending on the username, the user is directed to either the health care professional web page or patient web page. Moreover, a secure image capture was added to the login page to reduce brute force attacks. Text entered in the password field box is hidden and data transmitted between ages are hidden.

A patient welcome webpage was developed. This displays a search box for a table containing data logs, buttons for logging out, surgery details webpage, upload webpage and to continue to graphical webpage once a log is selected. Figure 4 reports an overall summary of the different webpages which are presented to the patient, a similar view is provided for the health care professional.

The upload webpage allows the user to upload files of data saved when importing data from the health monitor. The surgery details webpage, displays details of the surgery the patient is attending, this is shown in Figure 4.

When directed to the graphical webpage, the user can view all their sensor data in graphical form and are able to zoom into specific time points. Also, the user can go back to the previous page or go to webpages for a symptom checker or patient advice, this is shown in Figure 4.

The health care professional webpage, allows health care professionals to search for patients they have access to and view their data logs in graphical form.

3. Results

3.1. Secure Medical Wearable Device

The secure medical wearable device, shown in Figure 3, can measure vital body parameters and store the data with the date and time they are measured on the SD card. The communication with the device was successfully tested between the monitoring system and a personal computer; the data were also uploaded successfully to the database and accessed via the web application. The AES Bluetooth encryption algorithm properly encoded the data. The storage medium provides a mechanism for user authentication; therefore, data is stored in encrypted form before transmission via a public network or internet.

The device has been successful in these preliminary tests (i.e. sensing): the data are encrypted at the patients end before being transmitted to the secure database and the SD card is protected using authentication mechanisms.

3.2 Web Application

The successful implementation of the front-end web application is shown in Figure 4. This application is driven and populated using a secure relational database. The system allows for

multiple types of users with different security policies, for example a patient user and a medical professional.

The system is designed such that only the patient can upload, and view data associated with them. The medical professional can only access data that is associated with their patients. The graphs were able to show all the data recorded, allowing for a more accurate diagnosis.

To ensure a secure system many security methods were implemented, from a simply masking of the password and the one-way hashing of passwords which guards against the possibility that someone who gains unauthorised access to the database can retrieve the passwords of the users in the system. Also, used were functions that removed unauthorised characters to prevent SQL code injection type attacks. In addition, the system prevents brute force attaches by using a secure image capture facility. The system would also use transport layer security and security certificates, by hiding passwords entered, using hashed passwords to hide what has been typed in the field box, HTML special chars was used to prevent the webpages from being exploited with code injections and the secure image capture was added to reduce brute force attacks.

3.3 Ethical Issues with the System

As with any system that collects data about individuals there must be consideration of the ethical issues related to this system.

The collected data could be used against the patient by health insurances or wellness programs run by employers. This data could be shared with of health care clinics without the patient knowing. This data could be misused to harm the patient. This health monitor could be misused to keep track of people. For example, a business company or insurance company could force employees or customers to wear this device to keep track of them, instead of using the device for its purpose of wellness improvement and health awareness. It would be imperative that the system complies with all GDPR and that users of the system are trained and authorised.

4. Conclusion and Discussion

In this paper we presented a novel integrated system for the real-time monitoring of physiological parameters. The system provides an SD card reader for saving the data locally, encryption algorithms to encrypt the data, authentication services, and other security procedures. The wireless IEEE 802.15.1 secure communication protocol for transferring the data and a web service was used for the patient and the health care professional.

The system outlined in this paper was implemented successfully. It records physiological parameters, transfer data securely to a computer where it was stored in a secure database and was accessible by different users of different security policies (i.e. patient and health care professional).

The system automatically performs some analysis of the data and presents the information in graphical form. for both the patient and health care professional in graphical form using the web application.

Despite the preliminary performed tests, the system will have to be validated vs. golden standard instrumentations, namely medical and certified equipment [13-15]. Moreover, further laboratory trials should be performed to optimize the portability of the system and to make the interface more user-friendly in a daily medical and hospital scenario/environment. On this matter, an assessment on the reliability and accuracy of the system was not performed at this stage. Therefore, a preliminary validation of this system vs. golden standard instrumentation should be performed in order to validate the system. Such a validation may be based on comparing the measurements of the system with the same parameters as obtained from a polygraph or similar devices and by processing this data with a similar approach to [16, 17], namely Bland Altman technique or similar ones.

The device developed is a prototype and currently we are looking into ways to reduce the size of the energy source and increase the overall lifetime of the system before a recharge is required, as the current system will only last up to 10 minutes. We are also investigating a better, smaller design for the housing of the hardware, so that it is less invasive for the wearer. It is believed however, that point of concept has been proven.

In future these devices could be medically certified which would result in trust of the devices, the processes, and with the accuracy of the data.

Acknowledgment

This work was presented in thesis form in fulfilment of the requirements for the BSc in Computer Science and Electronic Engineering for the student Callum Owen-Bridge under the supervision of Dr. S. Blakeway and Dr. E.L. Secco from the Robotics Laboratory, Department of Mathematics & Computer Science, Liverpool Hope University

References

[1] World Health Organization. (2017). The top 10 causes of death. [online] Available at: [Accessed 6 Dec. 2017].

[2] Eisenach G. (2001). What is e-health?. Journal of Medical Internet Research, [online] Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1761894/ [Accessed 6 Dec. 2017].

[3] Alnosayan N, Lee E, Alluhaidan A, et al (2014). MyHeart: An intelligent mHealth home monitoring system supporting heart failure self-care. 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services.

[4] Government.nl. (2018). Benefits of eHealth. [online] Available at: https://www.government.nl/topics/ehealth/benefits-of-ehealth [Accessed 1 Mar. 2018].

[5] de Grood C, Raissi A, Kwon Y and Santana M (2016). Adoption of ehealth technology by physicians: a scoping review. Journal of Multidisciplinary Healthcare, 9, 335-344.

[6] World Health Organization. (2017). Global Diffusion of Ehealth. Geneva: World Health Organization.

[7] Virone G, Wood A, Selavo L, et al (2006). An Advanced Wireless Sensor Network for Health Monitoring. Transdisciplinary Conference on Distributed Diagnosis and Home Healthcare

[8] Li M, Lou W and Ren K (2010). Data security and privacy in wireless body area networks. IEEE Wireless Communications, 17(1), 51- 58..

[9] Omoogun M, Seeam P, Ramsurrun V, Bellekens X and Seeam A (2017). When eHealth meets the internet of things: Pervasive security and privacy challenges, 2017 International Conference on Cyber Security And Protection Of Digital Services.

[10] Meingast M, Roosta T and Sastry S (2006). Security and Privacy Issues with Health Care Information Technology. 2006 International Conference of the IEEE Engineering in Medicine and Biology Society.

[11] Cooking-hacks.com.(2018).e-HealthSensorPlatformV2.0forArduinoand Raspberry Pi [Biometric / Medical Applications]. [online] Available at: https://www.cooking-hacks.com/documentation/tutorials/ehealth-

biometric- sensor-platform-arduino-raspberry-pi-medical [Accessed 22 Mar. 2018].

[12] Bourdon R (2018). WampServer. Alter Way.

[13] Secco EL, Curone D et al (2012). Validation of Smart Garments for Physiological and Activity-Related Monitoring of Humans in Harsh Environment, American Journal of Biomedical Engineering, 2(4), 189-196

[14] Magenes G, Curone D, Caldani L, Secco EL (2010). Fire fighters and rescuers monitoring through wearable sensors: The ProeTEX project, Annual International Conference of the IEEE Engineering in Medicine and Biology Society

[15] Curone D, Secco EL, et al (2012). Assessment of sensing fire fighters uniforms for physiological parameter measurement in harsh environment, 16(3), 501-511, IEEE Transactions on Information Technology in Biomedicine

[16] Secco EL, Curone D, Tognetti A, Bonfiglio A, Magenes G (2012). Validation of Smart Garments for Physiological and Activity-Related Monitoring of Humans in Harsh Environment, American Journal of Biomedical Engineering, Vol.2, No.4, 189-196

[17] Curone D, Secco EL, Caldani L et al (2012). Assessment of Sensing Fire Fighters Uniforms for Physiological parameter Measurement in Harsh Environment, IEEE Trans on Information Technology in Biomedicine, vol. 16, no. 3, pp. 501-511

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 439-446 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Using Nullors to Modify Linear Model Parameters of Transistors in an Analog Circuit

Reza Hashemian*

Electrical Engineering, Northern Illinois University, DeKalb, IL 60115, USA

ARTICLE INFO

Article history:
Received: 30 June, 2018
Accepted: 11 October, 2018
Online: 20 October, 2018

Keywords:
Amplifier
Circuit analysis and design
Frequency responses
Model parameters
Nullors applications

ABSTRACT

Usually, after properly biasing an analog circuit we need to linearize its transistors in order to linearize the entire circuit for performance analysis and design. This linearization is done through modeling of devices, which is becoming more and more complicated as the electronic technology advances. This is not of course a major issue for circuit simulators that use the latest device models and fast computers to deal with them. However, it makes it difficult when manual computation and understanding of the circuit behavior is in question. This is particularly important in teaching and training electrical engineering students.

What the proposed method offers is to simplify the case in two steps. First, adopt a most recent model with nominal parameter values for the transistors and run the circuit. Second, compare the circuit responses with those obtained through simulators with the most advanced device models. Next, with the help of one or more nullors try to modify the manually selected model parameters so that the two responses become close enough together. Several examples demonstrate the way the technique works and how close we can get to the simulated responses.

1. Introduction

Linearizing analog circuits is a major step in circuit analysis and simulation, preparing for performance analysis and design. However, it is also important to understand the circuit behavior for design purposes. This is particularly essential in teaching students/designers in the field of electrical engineering. With so fast changing electronic technology and the transistors' modeling techniques, this linearization process becomes very cumbersome when it comes to hand calculations and manual circuit verifications and evaluations [1-3].

Several strategies can be proposed. A simple and effective one is to apply external probing such as experimenting with the transistors in different circuits and study their behavior, and then start building a working model. Alternatively, we can manually linearize the transistors to our best knowledge, and with an available and appropriate model that we think is working. Then we can compare responses from this circuit with those we get from an advances circuit simulator. This definitely will guide us to modify our model for a better accuracy. Here, in this study we are trying to focus on this proposed methodology. It is interesting to note that the proposed technique takes a full advantage of the computational power and the precision expected from a modern circuit simulator, in order to direct the manually modeling procedure to achieve its goals.

Here is how the method works. Let us first assume a circuit that has only one transistor. We choose a linear small signal equivalent circuit model for the transistor in the circuit that best represents the device modeling. This apparently linearizes the entire circuit and prepares it for the AC analysis. Our next step is to simulate the circuit. However, to be certain about the direction we take for the modification of the model elements, this simulation is done in combination with the circuit with the original transistor. The combined circuits here are connected in parallel-parallel format. Other formats are also possible, such as series-series, parallel-series, and series-parallel, depending on the types of the input and output ports. In this connection both circuits get the same input signal and will produce identical outputs. Naturally, considering circuit laws, this identical outputs cannot happen unless we allow an extra degree of freedom in the linearized circuit. To fulfill this task one component in the linearized model is allowed to freely change its variables (v and i) values, which basically becomes a

*Reza Hashemian, Electrical Engineering, Northern Illinois University, DeKalb, IL 60115, USA. Tel: 815-753-9930, rhashemian@niu.edu

norator. The final step is to find a closest real component (or any two-terminal circuit) that best suits to replace the norator.

In short, there are two main steps involved here. In the first step the analog circuit is manually linearized after the biasing is completed and the regions of the operations for the transistors are specified. This will allow us to roughly assign linearized model parameters to the devices and hence make the entire circuit linear. In the second step this linear circuit is linked to the actual nonlinear circuit, in its original form, and the combined circuit is simulated with the constraint that both circuits produce the same responses, without interfering each other's operations. This process can of course be repeated for as many times as needed, depending on the precision required. In each process a different model component is selected for change, or alternatively, new components can be added to the model until we are satisfied.

Overall, using this procedure gets us closer and closer to the ideal response for the application we have. In exchange the price we need to pay is to modify one or more device model components in a manner in which the output response will improve to the ideal case even when we disconnect the two circuits from each other. This means, the linear circuit responds favorably after being separated from the original circuit. It must also be noted that this similarity in responses must only be valid for a certain region of operation, and for a desired bandwidth. That is, depending on the selections of the model components, the situation only needs to work for the application in hand and the frequency bandwidth described. Outside of this region of operation or bandwidth the response may or may not match with that of the original circuit.

The crucial components that are used for such operations are nullors, which are special types of fixator-norator pairs (FNP) [4-10]. Nullors are among pathological elements that although not ideally available in physical form but have very unique properties particularly for circuit designs [8, 9]. Briefly, a nullor consists of a pair of elements, a nullator and a norator. A nullator element passes zero current and has zero voltage across, whereas, a norator element has both its current and voltage unspecified. The difference between a nullor and an FNP is in the nullator element. In case the nullator in a nullor allows to have any arbitrary (but fixed) current or voltage then the nullor becomes an FNP. When used in pairs the nullator keeps both component variables (v and i) fixed, whereas both variables of the pairing norator are specified by the circuit. It is this unique property that makes a nullor well suited for design purposes. And here we exactly use the pair to adjust and modify the transistor model parameters in order to make the linearized circuit to perform as specified.

Finally, it must be pointed out that there is no claim this methodology can replace a formal and accurate transistor modeling, which is typically used by circuit simulators. The main purpose of using this technique is to educate electrical engineering students to manually experiment with device modeling for different applications. In addition, the approach simplifies and makes it more understandable to study circuit behavior such as finding poles and zeros and frequency behavior of the circuit. Although not still to its perfection, the method seems to be new and practical. As far as this author is aware of, there is no similar method reported in the literature to compare.

2. Modification of Model Parameters in a Linearized Circuit

The objective in this presentation is to linearize an analog circuit with transistors after being biased for a certain application. This is needed to perform further analysis such as AC analysis, transient analysis, and also for performance design [11]. A typical performance design procedure for amplifiers and analog filters is to design for frequency profile and bandwidth. To do this we need to first linearize the devices in the circuit and come up with a linear circuit. The next step is to perform AC analysis on the circuit using a circuit simulator. The procedure is done by first replacing the transistors in the circuit with their small signal linear models that are constructed from the device characterization and parameters extractions. Depending on several factors, such as the fabrication technology used, device sizes, and the transistors' modeling levels (in Spice) being adopted, these model parameters can become very complicated for manual considerations. They may even change whether we use them for analog design or digital. In general getting quite accurate model parameters mainly for short channel MOS transistors, for example, is not a trivial task. Figures 1 and 2 are examples of simpler linearized models for longer channel lengths MOS transistors.

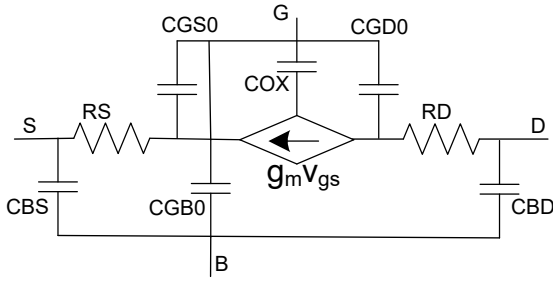


Figure 1. SPICE Level 1 Model for long channel MOS transistors.

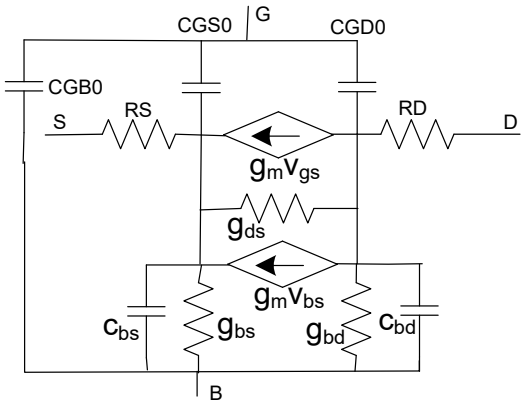


Figure2. SPICE Level 3 Model for long channel MOS transistors.

Figures 1 is the Level 1 model for long channel MOS transistors adopted by SPICE simulation program, and Fig. 2 is the schematic of a more advanced, Level 3, model with the code given below [12].

```
*Model parameters for long channel nMOS Level=3 models, supply VDD=5V
.MODEL N_1u NMOS LEVEL  = 3
```

```
+ TOX   = 200E-10       NSUB   = 1E17         GAMMA  = 0.5
+ PHI   = 0.7           VTO    = 0.8          DELTA  = 3.0
+ UO    = 650           ETA    = 3.0E-6       THETA  = 0.1
+ KP    = 120E-6        VMAX   = 1E5          KAPPA  = 0.3
+ RSH   = 0             NFS    = 1E12         TPG    = 1
+ XJ    = 500E-9        LD     = 100E-9
+ CGDO  = 200E-12       CGSO   = 200E-12      CGBO   = 1E-10
+ CJ    = 400E-6        PB     = 1            MJ     = 0.5
+ CJSW  = 300E-12       MJSW   = 0.5
```

For short channel devices BSIM models are well advanced that are developed by Berkeley. In comparison with the lower level models, BSIM models are covering sub-micron and way down into nanotechnology cases. An example of 50nm BSIM4 models with V_{DD}=1V is a page long, and it is partially given below [12].

```
* Model parameters for 50nm nMOS BSIM4, supply VDD=1V.
.model  N_50n  nmos  level = 54
+binunit = 1         paramchk= 1         mobmod  = 0
+capmod  = 2         igcmod  = 1         igbmod  = 1        geomod  = 0
+diomod  = 1         rdsmod  = 0         rbodymod= 1        rgatemod= 1
+permod  = 1         acnqsmod= 0         trnqsmod= 0
+tnom    = 27        toxe    = 1.4e-009  toxp    = 7e-010   toxm    = 1.4e-009
. . . .
. . . .
. . . .
+dmcg    = 0e-006    dmci    = 0e-006  dmdg  = 0e-006   dmcgt   = 0e-007
+dwj     = 0e-008    xgw     = 0e-007    xgl     = 0e-008
+rshg    = 0.4       gbmin   = 1e-010    rbpb    = 5        rbpd    = 15
+rbps    = 15        rbdb    = 15        rbsb    = 15       ngcon   = 1
```

As we can see, putting a model BSIM4 into schematic for manual calculation/evaluation is not practical and almost impossible. Although a circuit simulator, such as SPICE, does actually replace MOS transistors with such models for AC analysis, but doing so manually and for educational and training purposes is not simple, and even some of the model parameters cannot be translated into physical entities. So, the model must be simplified. One possibility is to do it for each case separately, and based on the circuit/device behavior. It may take a bit longer but it can be worth understanding the device behavior in that particular region of operation and for that application.

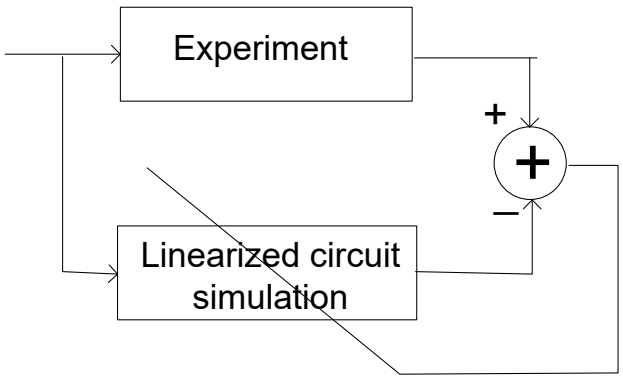


Figure3. Modification of the linearized transistor model parameters is directed by experiments or in comparison with the simulation of the original circuit.

In our proposed methodology the process is done in two steps. First, the circuit is manually linearized by replacing the transistors with their linear model parameters that are best available for the region of operation and the bandwidth specified. We next simulate the linearized circuit as well as the original analog circuit and then compare the two results. If the responses happened to be close enough then we have hit the target, and the model parameters have been rightly selected. Otherwise, we go through the second step, in which we select one of the model elements in the circuit that seems most effective in changing the output response. Then we will proceed into a procedure that tries to modify the model element aiming at a response that mimics the response from the original circuit.

All we are trying to do in this methodology is to focus on this second step of the process, and find the right model component for the transistor. In case one selected model element does not finish the job to our satisfaction, we can still follow the same procedure with another model component. This is well illustrated in Fig. 3. As shown, the experimental results getting from a spectrum analyzers or a digital oscilloscope are compared with the results coming from simulating the linearized circuit. Alternatively, both the original circuit and the linearized one can be simulated in combination. Now, if the two responses are still too far apart the difference is fed back into the system to further modify the selected model component in the linear circuit. This is a sort of an ad hoc adaptation procedure. The question is, what is the purpose of manually linearizing an analog circuit when we don't have the exact model parameters for its transistors? Or what are the advantages of linearizing a circuit when we do not have access to the accurate model parameters? The answer to these questions mainly rely on the simplicity and the analytical power that exits with linear circuits. The experiments on nonlinear circuits are limited to the conditions and regions that the circuit operates in. After the biasing of the transistors in a nonlinear setting are completed we need to do performance (AC) design. We need to linearize the devices using model parameters that are valid for the selected regions of operations and the bandwidth we need.

So, circuit linearization shifts our circuit analysis and design from nonlinearity into linear domain, with all its properties and advantages. Even if a set of manually selected linearized model parameters are properly working for a certain operating condition then we can claim that we have reached to a regional situation, and in case the operating conditions of the transistors shift to a new region we can still try to modify those parameters again to reach to a solution for the new case. After all, we don't forget that this is just a simulation and the purpose is to explain the devices behavior in certain regions of operations, and in an educational setting.

Another point worth mentioning here is that, depending on the operational environment, it is most preferred to compare the results of a manually linearized circuit with experiments that are done with the real circuit. This is of course more easily done in case of a laboratory environment, but for a class or study room setting where only simulation facilities are available the case is different. Hence, for the latter case, it is more practical to have both original circuit, as the reference circuit, and the linearized circuit combined in one circuit and then simulate.

2.1. Linear model parameter identification

The type of arrangement just described is illustrated in Fig. 4 for a single model parameter change. Figure 4(a) is a symbolic representation of the linearized circuit along with a selected model element, denoted by P, to be modified. In Fig. 4 (b) the actual

circuit, denoted by *M*, is connected in parallel with the linearized circuit *N* ready for simulation. The responses of the circuits, i.e., the linearized *N* and the original circuit *M,* are then continuously compared. To keep the two responses identical, a nullator is used in the output port. Notice that the nullator, in addition of keeping the two output responses identical, it also provides independence for each circuit to operate without being interfered by the other. In exchange, it is the pairing norator that is responsible to provide the condition needed for such situation.

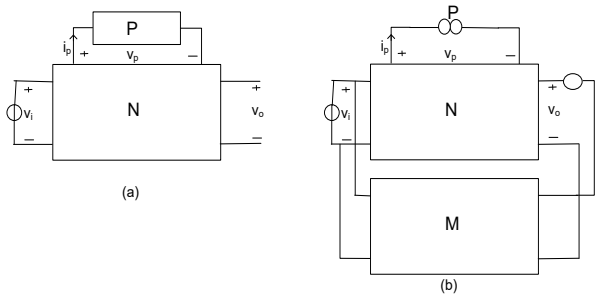


Figure4. (a) Manually linearized circuit *N* with a variable element *P* to be added; (b) nullor is added to be able to realize the element *P*.

Now, there are two cases that might happen. Either the v – i characteristic of the norator closely matches with that of the selected model component or they are way apart. For the former case the experiment is over and the model component remains unchanged for the linearized circuit.

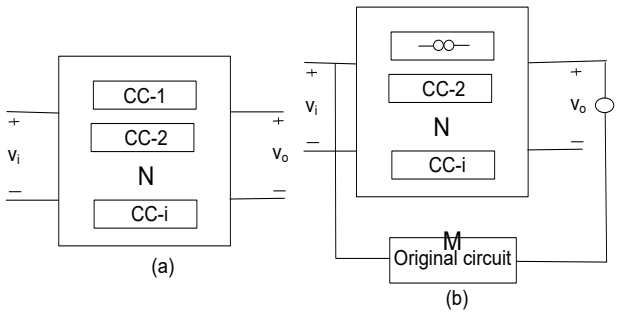


Figure 5. (a) Manually linearized circuit *N* with one or more possible sub-circuits *CC*; (b) nullor is added to keep the interconnection between the linear circuit and the original circuit ordered.

In the second case, we find that the model component and the norator do not match, which means we need to modify the model component so that its v – i characteristic is getting close enough to that of the norator. However, it is always possible that a single model parameter (component) may not completely solve the problem. In this situation we need to select another model component and continue with the same process, and so on. This situation is displayed in Fig. 5 with multiple Circuit Components (CCs) to be considered. Again, Fig. 5(a) shows the linearized circuit with *i* number of model elements, denoted by CC, that are candidates for changes. In Fig. 5 (b), on the other hand, the actual circuit, denoted by M, is added and connected in parallel with the linearized circuit. The rest of the testing procedure is the same as was explained previously. The difference, however, is that here we go through simulations multiple number of times and in each time one of the CC components is replaced with a norator. This certainly continues improving the performance of the linearized circuit in reference to the actual circuit, in multiple steps until we get satisfied with the results.

Now we reach to a point that we can put the entire process into an algorithm for implementation.

Algorithm 1 – Consider a circuit *M* with transistors. Let *M* be manually linearized and call the resulted circuit *N,* as described earlier. We can then go through a stepwise procedure that modifies and corrects one or more model elements of the transistors in order to make the response of the linearized circuit to closely follow the response given by the original (nonlinear) circuit. The rest of the procedure is as follows:

1. Search for one or more transistor small signal model elements in *N* and call them CC_j, for all j, as shown in Figs. 4(a) and 5(a). Connect *N* and *M* circuits together through a nullator. For now make this connection parallel-parallel (for other possibilities see Footnote 1), as shown in Figs. 4(b) and 5(b).
2. Chose one of the model components *CCs*, say CC_j. This is a candidate to be modified for the design. Replace CC_j with the paring norator.
3. Perform a simulation on the combined circuit, *M* and *N*, for AC analysis and for the bandwidth that is specified.
4. Assume $v_j(s)$ and $i_j(s)$ to be the voltage across and the current through the norator CC_j, respectively. Then find the equivalent pseudo-impedance $z_j(s) = v_j(s)/i_j(s)$ of the norator, and plot its Bode plot p_j for the bandwidth given.
5. Next, try to replace the norator CC_j with an actual component, or a two-terminal circuit, that ideally has the same impedance $z_j(s) = v_j(s)/i_j(s)$. In case such an ideal case is not reached, find the closest component (or a two-terminal circuit) CC_j so that its Bode plot response p'_j is close enough to p_j, for the given bandwidth.
6. Replace the norator (CC_j) with the component (or the two-terminal circuit) CC_j found. Next, disconnect *M* from *N* circuit, and then simulate *N* alone for its response.
7. If more modifications are still needed, go to step 2 and select a different model component, say CC_k in *N* and follow the rest of the procedure. Continue with the procedure, as many times as needed, until an acceptable response is obtained.

3. Examples

Several examples are worked out here that illustrate the way the proposed method works. Example 1 assumes a circuit with a single nMOS transistor. Instead of using the exact model parameters, given in the data sheets, we simply use level 3 Spice model for long channel devices with values most appropriate. Next, we try to modify the parameter values until we get close enough to the desired response. To do this we use the FNP technique, described before, to adjust the model parameters within the bandwidth specified. Example 2 is given for a two stage nMOS amplifier with multiple transistors, where the type and size of the transistors are the same as those in Example 1. Therefore, we should be able to utilize the same transistor model parameters developed for Example 1. Example 3 is yet another nMOS transistor amplifier that utilizes FNP technique to modify the transistor model parameters. Here also we keep our previous level 3 transistor model used in Examples 1 and 2 unchanged. The point

of repeating the same model for all three cases is to see how expanding the circuits with more transistors but with the same transistor linear modeling affects the accuracy and the validity of the proposed method.

Example 1: – Take a single stage MOS amplifier, as shown in Fig. 6(a). Next, proceed to linearize the amplifier by using level 3 model parameters for its transistor. The assumption is that we

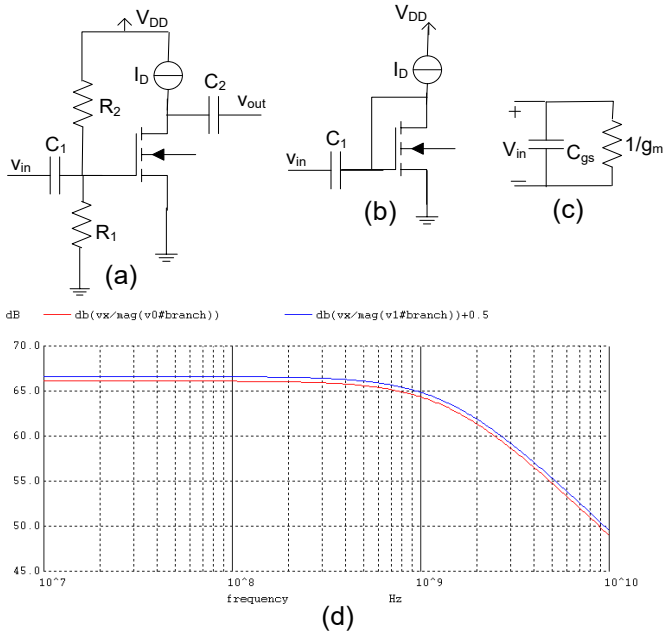


Figure 6. (a) An nMOS transistor amplifier; (b) diode connection; (c) equivalent diode connection in the linearized amplifier; (d) comparing plots for the input impedances in (b) and (c).

have no access to the Spice parameter values at this point, although the simulator is assumed to use the latest device model. To get the pseudo-parameters for our linearized model we go into two operational steps. First, we connect the transistor in a diode format (in the original circuit as well as in the linearized circuit) and then simulate the two circuits combined. Figures 6(b) and (c) are the two circuit schematics, and the Bode plots resulted from the two impedances, looking through the input ports, are demonstrated in Fig. 6(d). Note that for readability purposes one plot is artificially raised by 0.5 dB, otherwise the two are exactly the same.

Now, in the next step we use an FNP (similar to nullor) procedure to find the remaining parts of the model parameters for the transistor. To do that, we can simply use the actual amplifier (Fig. 6(a)) as the "model" circuit to help to adjust the model parameter in the linearized circuit, as given in Fig. 7(a). The Spice circuit code for the combined circuit is also given below, where a voltage controlled voltage source (e1) is used to represent an FNP.

```
.control
destroy all
set units = degrees
ac dec 1000 1.0e6 1.0e11
plot db(v(3)) db(v(6))+1
.endc
******** Combined circuits *******
Vin     2       0       DC      0       AC      1
x11     2       3       non-Lin
x21     6       2       0       Lin-Mos
********** Circit 2 *********
.subckt Lin-Mos  7       6       0
gm1     7       0       6       0       0.495m
cgs1    6       0       55.8f
*VCVS   e1      6       7       3       7       1.0e06
rgd1    6       7       2.2856MEG
cgd1    6       7       9.7f
.ends
***************** Ciruit 1 ******
.subckt non-Lin  3       7
VDD 2   0       DC      5
id1     2       5       20u
M11     5       4       0       0       N_1u L=1u W=50u
r11     4       0       198k
r21     2       4       1MEG
c11     3       4       10u
c21     5       7       10u
.ends
.include cmosedu_models.txt
.end
```

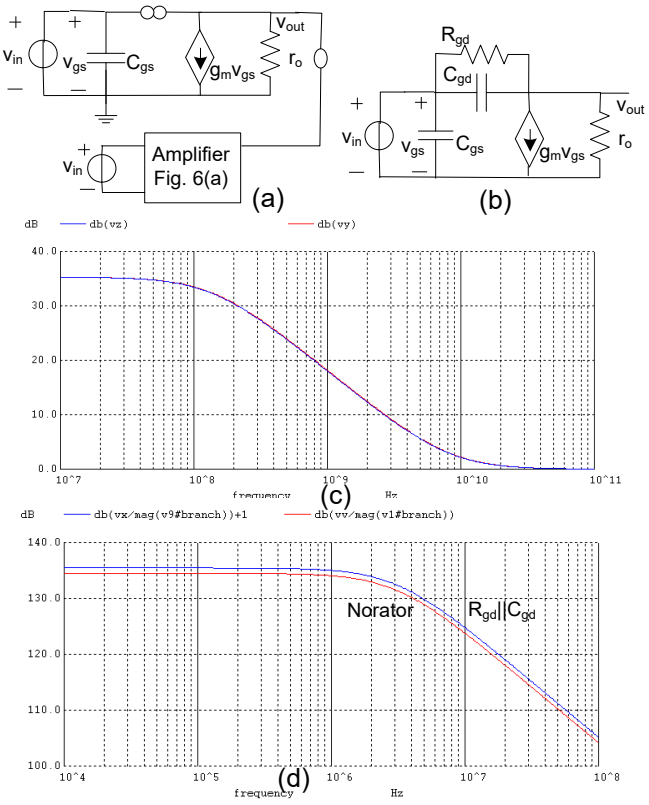


Fig.7. Getting the model parameters for the nMOS; (a) assumsed linearized circuit is parallel with the amplifier circuit; (b) nMOS linear model circuit; (c) gain responses for the actual amplifier and its linear equivalent circuit; (d) plots for the impedance functions, the norator and the replacement with the parallel R_{gd} and C_{gd} circuit.

In Fig. 7(c) we notice one gain plot from the actual amplifier and one from the linearized circuit. Obviously the two must be identical because a nullator enforce them to. Next, we need to replace the norator with one or more real component. To find this component we first run the circuit with the norator and make a plot of the impedance function of the norator. This plot is given in Fig. 7(d). The ideal solution is possible if we could find one or more combination of real components that represent the same impedance plot. Here we notice the norator plot is in fact very close to the impedance plot of a parallel RC circuit. This RC

circuit is represented by R_{gd} and C_{gd} in the linearized circuit given in Fig. 7(b), and the impedance function of the $R_{gd} \parallel C_{gd}$ is also plotted in Fig. 7(d), which is almost identical to the norator impedance (except for the 1dB clearance). This concludes our Example 1.

Example 2: – Consider a two stage nMOS shunt - shunt feedback amplifier, as shown in Fig. 8 [12]. There are four identical nMOS transistors used in this amplifier. In order to minimize our modeling efforts, we are going to use the same type and size transistor we used in Example 1. With this choice, to linearize the amplifier circuit all we need to do is to replace each transistor with its linear model already found in Example 1 with a bit of adjustment. This adjustment is done on the resistance r_{gd}, which is in parallel with c_{gd} in the transistor model. The value of r_{gd} in Example 1 was found to be close to 2.3 MEG ohms, whereas in Example 2 it was increased to 40 MEG ohms, where both resistances are quite large anyway. The list of the Spice program for this example is given below.

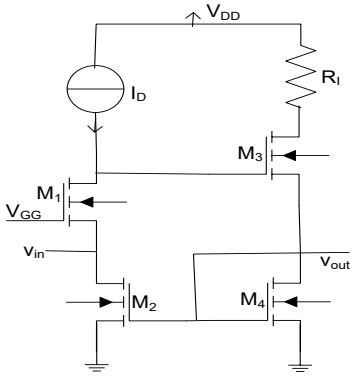


Figure 8. A two stage nMOS current amplifier.

```
.control
destroy all
set units = degrees
ac dec 1000 1.0e5 1.0e10
plot db(I(v21)) db(I(v31))
plot ph(I(v21)) ph(I(v31))
.endc
****** The combined circuit ******
ii1     0       2       DC      0       AC      1
x11     2       3       Mos-2
x21     2       4       M-Linear
v21     3       0       DC      0
v31     4       0       DC      0
****** Linear circuit ********
.subckt         M-Linear  8       9
x11     6       0       3       Lin-Mos
x21     3       4       0       Lin-Mos
x31     7       6       4       Lin-Mos
x41     4       4       0       Lin-Mos
r11     0       6       160k
r21     0       7       100k
c11     3       8       10u
c21     4       9       10u
.ends
*************** Linear-Modeling ********
.subckt         Lin-Mos 7       6       3
gm1     7       3       6       3       0.52m
cgs1    6       3       55.8f
rgd1    6       7       40MEG
cgd1    6       7       9.7f
.ends
***************** The original Amplifier ********
.subckt         Mos-2   8       9
VDD1    2       0       DC      5
VG1     5       0       DC      1.770833
M11     6       5       3       0       N_1u  L=1u W=50u
M21     3       4       0       0       N_1u  L=1u W=50u
M31     7       6       4       0       N_1u  L=1u W=50u
M41     4       4       0       0       N_1u  L=1u W=50u
r11     2       6       160k
r21     2       7       100k
c11     3       8       10u
c21     4       9       10u
.ends
.include cmosedu_models.txt
.end
```

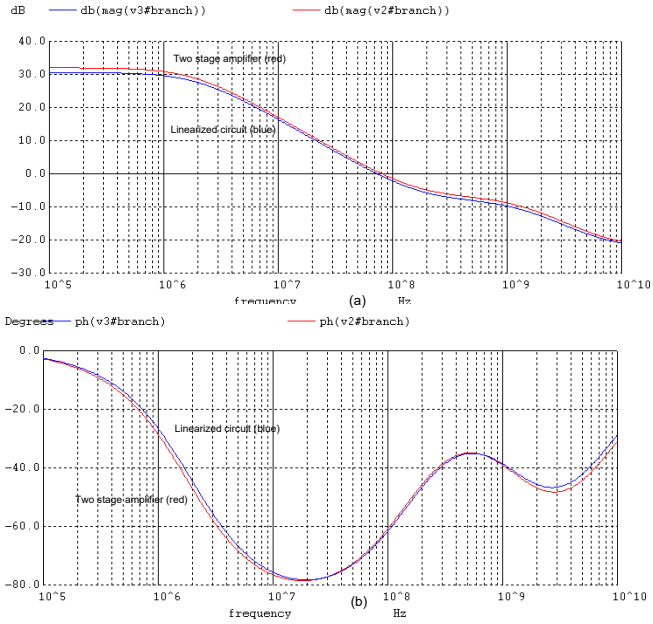


Figure 9. Bode plots given for the gains for the two stage nMOS amplifier and its linearized circuit; (a) magnitude plots; (b) phase plots.

Now it is time to simulate the amplifier and its linearized equivalent circuit in combination. The simulation generates the output transfer functions plots given in Figs. 9(a) and (b), where the plots in (a) represent the magnitude and in (b) the phase angle plots. Notice that the plots are almost identical for the original amplifier and its linearized circuit. Further experiments show that with the body effect removed in this example the two responses become even closer and tighter together.

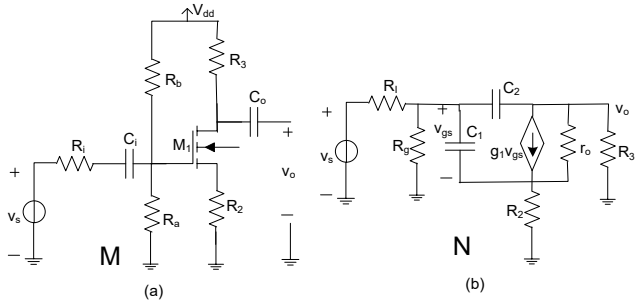


Figure 10. nMOS amplifier circuit and its linearized equivalence. (a) The original nMOS amplifier, and (b) its AC small signal linearized equivalent circuit.

Example 3: – Consider a single stage nMOS amplifier *M* with

feedback, as shown in Fig. 10(a). As usual, we first replace the transistor with its small signal model suitable for this case. This linearizes the amplifier and ready to compare its transfer function with the original amplifier. The model parameters used in this example are assumed to be the Spice Level 3 parameters provided for long channel nMOS transistors. Figure 10(b) shown the amplifier circuit after being linearized.

Our next move it to simulate both circuit, *N* and *M*, and plot the transfer functions. Figures 11 represents the plots for (a) the magnitude, and (b) the phase. In comparing the results from the two circuits we notice the differences, particularly in the magnitude plots; although the plots show the same patterns of variations vs the frequency and also in term of poles and zeros locations.

Now the question is what has gone wrong here, and how can this be corrected? In our efforts to find out about this discrepancy we realize that the ratio between the values of the two transistor model capacitors, i. e., $C_{gs} = 2.05\ pF$ and $C_{gd} = 13.83\ fF$, is very high and close to 150. So, C_{gs} seems to be too high and must be modified. We pick up C_{gs} and replace it with a norator, and go through Algorithm 1 for the adjustments.

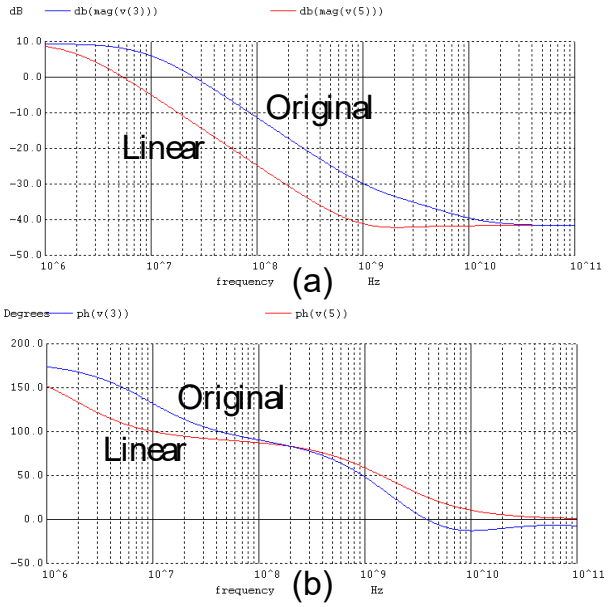


Figure 11. Transfer function responses given for the original amplifier circuit and it linearized equivalent circuit. The differences shown in the responses are mainly the results of mismatch in model parameters. (a) The magnitude, and (b) the phase.

As stated in Algorithm 1, step 2, we are going to make a parallel-parallel connection between the original circuit, *M,* and the linearized one, *N,* through a nullor (this connection is the same as that in Fig. 7(a)) and is not shown here. For the norator we choose a current controlled voltage source (CCVS). Next, we use Spice simulation tool to simulate the combined circuit. The partial Spice programming code for this example is given as.

```
.control
destroy all
let vm1 = v(5)-v(7)
plot db(v(6)) db(v(4))
plot db((vm1)/I(v11))  db((v(8))/I(v31))
.endc
.option scale = 1u
.ac dec 1000 1Meg 100G
VDD1  2     0     DC    5
Vs1   3     0     DC    0     AC    1
*************** The actual amplifier *****************
.subckt     model-1  2     3     6
M11   8     5     7     0     NMOS L=2 W=100
ra1   5     0     200k
rb1   2     5     330k
ri1   5     3     100k
r31   2     8     5.1k
r21   7     0     1k
ci1   5     4     10u
co1   8     6     10u
.ends
************** The linearized circuit ****************
ri1   3     5     100k
r21   7     0     1k
rg1   5     0     82k
r31   8     0     5.1k
ro1   8     7     180k
c11   5     7     2.05p
c21   5     8     13.83f
g11   8     7     5     7     1.648m
*********** Circuits combined ****************
x11   2     3     4     model-1
v01   7     4     DC    0
v11   5     b     DC    0
h11   b     7     v0    1.0e8
* *********** Testing Bench ***************
v31   3     9     DC    0
rx1   9     a     190
c31   a     0     300f
.include cmosedu_models.txt
.end
```

We notice from the simulation results that in order for the linearized circuit to respond exactly similar to that of the original amplifier is the followings. We must replace the norator with one or more components that in combination represent an impedance function with the characteristic plot given in Fig. 12(a), plot *CC.* Now, with a close look at this plot we come to a conclusion that this plot cannot be just from a capacitor C_{gs} alone. There must be a series combination of a capacitor and a resistor. This is because we see at least one zero on the real axis in the LHP. This *RC* circuit is shown in Fig. 12 (b), where we find $C = 300\ fF$ and $R = 190\ \Omega$. Next, when we plot the impedance function of the RC circuit we come up with the plot *RC* given in Fig 12(a). As noticed, the two plots, RC and CC, are now so close together, which are almost undistinguishable from each other. Our next move is to replace the norator in N with the RC circuit just found.

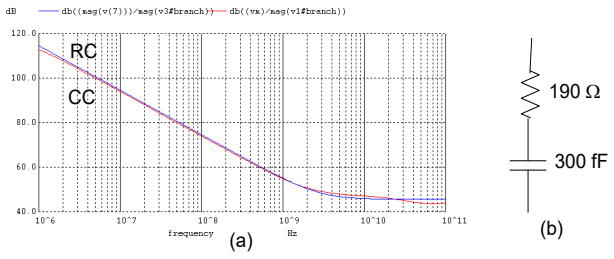


Figure 12. (a) Comparing Bode plots associated with two impedances; one representing the norator (CC), and the other one the RC circuit. (b) The series RC circuit.

After we have made the substitution (i. e., the original model element $C_{gs} = 2.05\ pF$, is replaced by the components $C = 300\ fF$

and $R = 190\ \Omega$ in series) we simulate the circuits, this time disconnected from each other, except for the source signal. The frequency responses of both circuits are now plotted in Fig. 13, (a) for magnitude, and (b) for the phase. Comparing the plots in Figs. 13 (a) and (b) with those in Figs. 11 (a) and (b) we see drastic changes. There are much more improvements, and this is particularly evident from the magnitude plots given in Fig. 13(a).

It is of course possible to go one step further and pick up another component in the transistor model for change. The process will be the same as explained before, i. e., replacing the component with a norator first and then follow Algorithm 1 for further processing. The result is expected to get the linearized circuit response closer to that of the original amplifier. However, we think the error is negligible at this point and there is no need to go further.

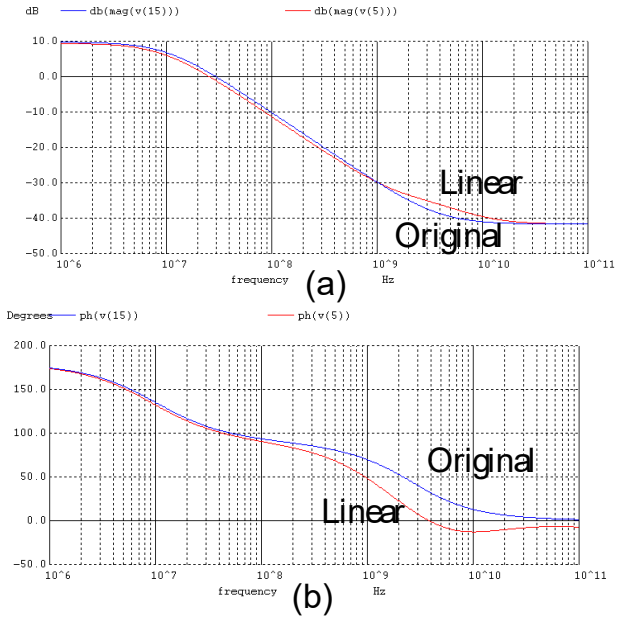


Figure 13. Comparing two transfer functions plots, one associated to the original nonlinear amplifier, and the other one to the linearized equivalent circuit; (a) magnitude, and (b) phase responses.

4. Conclusion

A simple and practical techniques is presented to manually linearize an analog circuit. The method starts with linearizing the transistors based of simple and available models but it gradually modifies one or more model components until an acceptable result is reached.

The method simplifies the case in two steps. First, adopt a most recent model with nominal parameter values for the transistors and simulate the circuit. Second, compare the circuit responses with those getting from the original circuits when simulated or in experiments. Then with the help of one or more nullors modify the manually selected model parameters so that the two responses become close enough for the application. Several examples demonstrate the way the technique works.

The method is especially useful for teaching electronic circuits to students and trainees, where understanding of how linearized modeling of devices works is of prime interest. It is shown that this modeling procedure is completely external to the device and works based on the circuit application and for an assigned bandwidth.

In this procedure the circuit is first manually linearized, which means replacing the transistor with a linear model that is available to the designer. The next step is to combine the linearized circuit with the original circuit in such a way that they behave independently without any interfering each other, but at the same time their responses remains identical. From circuit theory stand point, this is only possible if we allow a component (norator) in the linearized circuit to have varying (v and i) variables. The next step in the process is to replace the norator with a real component or sub-circuit in the transistor model. Evidently, this may not always work quite accurately due to the realizability of the component. In case the result obtained after this substitution is still unacceptable the process can go on including other components of the linearize model. That is, the process of modifying and correcting the manual modeling of the device can continue until the final response we receive is working and acceptable.

References

[1] T.L. Pillage, R.A. Rohrer, C. Visweswariah, Electronic Circuit & System Simulation Methods, McGraw-Hill 1995.
[2] A. S. Sedra, and K. C. Smith, Microelectronic Circuits, Seventh Edition, Oxford University Press, 2015.
[3] E. Tlelo-Cuautle, Advances in Analog Circuits, InTech, 2011.
[4] C. Sánchez-López, A. Ruiz_Pastor, R. Ochoa-Montiel, and M. A. Carrasco-Aguilar, "Symbolic Nodal Analysis of Analog Circuits with Modern Multiport Functional Blocks," Radio Eng., vol. 22, no. 2, pp.518-525, June 2013.
[5] E. Tlelo-Cuautle, C. Sa´nchez-Lo´pez, E. Martı´nez-Romero, Sheldon X.-D. Tan, "Symbolic analysis of analog circuits containing voltage mirrors and current mirrors," Analog Integr Circ Sig Process 65:89–95, DOI 10.1007/s10470-010-9455-y, 2010.
[6] M. Fakhfakh, E. Tlelo-Cuautle, F. V. Fernández, "Design of Analog Circuits Through Symbolic Analysis," *https://books.google.com/books?isbn=1608050955, Computers 2012.*
[7] C. Sánchez-López *et al*, "Pathological Element-Based Active Device Models and Their Application to Symbolic Analysis," IEEE Trans. Circuits Syst. I: Regular Papers, vol. 58, no. 6, pp 1382 – 1395, June 2011.
[8] R. Hashemian, "Application of Fixator-Norator Pairs in Designing Active Loads and Current Mirrors in Analog Integrated Circuits," IEEE Trans. Very Large Scale Int. (VLSI) Syst., vol. 20, no. 12, pp 2220 – 2231, Dec. 2012.
[9] R. Hashemian, " Fixator-Norator Pairs vs Direct Analytical Tools in Performing Analog Circuit Designs," IEEE Trans. Circuits Syst. II, Exp. Briefs, vol. 61, no. 8, pp. 569 - 573, August 2014.
[10] R. Hashemian, "Application of Nullors in Designing Analog Circuits for Bandwidth", IEEE Inter. Conf. on Electro/Information Tech., EIT2016 conference proceedings, Grand Forks, ND, May 19 - 21, 2016.
[11] R. Hashemian, "Nullors in Finding/Modifying Transistor Parameters, Blindly", IEEE International Conference on Information and Computer Technologies (ICICT), DeKalb, Illinois, March 23-25, 2018.
[12] R. J. Baker, CMOS, Circuit Design, Layout, and Simulation, John Wiley &Sons, New Jersey, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 447-453 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Economic and Social Sustainability for Iraqi Middle Provinces

Maysoon Abdullah Mansor*

College of Engineering, Dept. of Civil Engineering, University of Tikrit, Iraq

ARTICLE INFO

Article history:
Received: 13 August, 2018
Accepted: 16 October, 2018
Online: 25 October, 2018

Keywords:
Economic sustainability
Social sustainability
Sustainability Index

ABSTRACT

This research is part of a study concerned with building an integrated framework for development of sustainable urban planning in Iraqi provinces. Seven Iraqi middle provinces are covered in this study. The aim of the study is to determine, measure and evaluate economic and social sustainability in the Iraqi middle provinces.

The stages to achieve this research were based on establishing a two levels hierarchy framework (a level for categories and a level for indicators). In this study, the Economic dimension comprised of 6 categories and 18 indicators and the Social dimension comprised of 8 categories and 41 indicators. Through using the Multi Criteria Analysis method and the Analytic Hierarchy Process, the Sustainability Index has been determined and used to measure and evaluate the economic and social sustainability in the Iraqi middle provinces

1. Introduction

More than 50% of the worldwide population live in urban or urbanizing areas. In just five years, we will add another people, all will need jobs, food, security and opportunity. Sustainable development have become imperative to meet their own needs. [1], [2]

Generally , sustainable development concept comprises of three aspects: economic, environmental and social. There are other sources that show other aspects: sustainability, aaccording to LA21, has four dimensions, namely: economic, environmental, social and governance or institutional aspects [3], shows that the three sustainability main dimensions, all located within the governance dimension that manages and controls these dimensions.

1.1. The economic dimension

It is considered as a base for development and any disruption of it or to the draining of its resources eventually will lead to the weakening of its future development opportunities, and then a long-term economic perspective must be taken to solve the problems in order to provide effort, money and resources [4].

1.2. The social dimension:

It is the right of a normal person in a clean and sound environment in which he can practice all activities while ensuring his right in a fair share of natural resources and environmental and social services, and investing it in order to serve his basic needs (shelter, food, clothing, air ....) as well as the complementary needs to raise his standard of living (business, entertainment, fuel ....) and without reducing the chances of future generations[4].

1.3. Indexes

An index is a synthesis of indicators or composite indicator. Formalization of an indicator that necessitates the aggregation of several data or variables results in. In the field of sustainable development, the use of indexes facilitates the interpretation and understanding of given phenomenon indicators, particularly for the public. Indexing offers an instrument for benchmarking performance and provides information to enable decision-making and directing development [5]. According to its specific goals, Indicators should tell us in what fields the city is doing better than in others, and a single sustainability index should tell us whether the city is becoming more sustainable or not. Based on an aggregation method, the index is then calculated [6].

Composite indicators development is considered to be a unique approach for sustainable development evaluation. Calculating aggregated values is one of the common methods for establishing indices. Depending on its application, composites indices can be established without or with weights [7].

2. Methods Used in the Developing Sustainability Index

For developing mathematical sustainability index, the Methods used are:

*Maysoon Abdullah Mansor, Email: maysoonm810@gmail.com dr.maysoonabdullah@tu.edu.iq. This paper is an extension of work originally presented in 2018 1st International Scientific Conference of Engineering Sciences - 3rd Scientific Conference of Engineering Science.

2.1 The Analytic Hierarchy Process (AHP)

This method is most widely used Multi Criteria Decision Making method, as it presented a deriving priorities and structuring the decision problem logically and representatively. The use of pair-wise comparisons, in AHP method is, to compare the alternatives regarding to the various criteria and to assess weights of criteria" [8].

The comparisons and priority calculation of AHP pairwise are executed to get the relative weights for objectives, sub objectives, and measures on every level of criteria in the hierarchy [9]. Figure 1 describes the principal steps for achieve the AHP. Table (1) shows a preference scale using for comparison. Ref. [10] suggested using the following consistency index (CI) as in (1):

$$CI=\frac{\lambda_{max}-n}{n-1} \quad (1)$$

Where

CI: is the consistency index

λ_{max}: is the largest Eigen value of matrix and

n: is the number of elements within a branch being compared.

This consistency index can also be expressed as a consistency ratio as in (2):

$$CR = \frac{CI}{RI} \quad (2)$$

where:

CR: is the consistency ratio

RI: is a known random consistency index obtained from a large number of simulation runs and varies depending upon the order of matrix.

Table 2 shows the random consistency index values (RI) for matrices of order 1 to 10 obtained by approximating random indices using a sample size of 500[11].

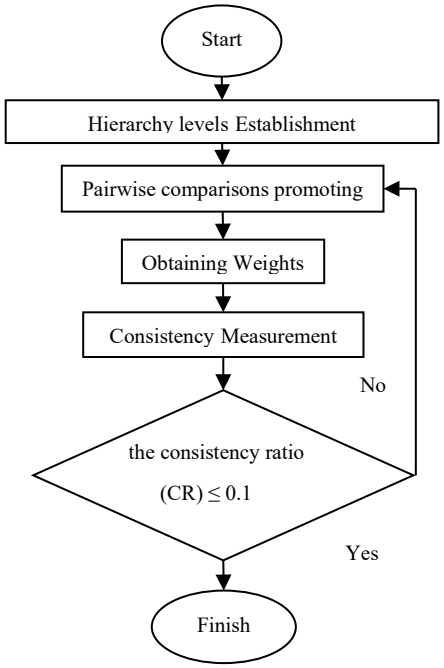


Figure 1 The principal steps for achieve the AHP [Researcher depending on [9]]

2.2 The Geometric Mean [12]

Equation (3) or (4) is used to find geometric mean:

$$G = \sum_{f_i}\sqrt{\left(x_1^{f_1}\right)\left(x_2^{f_2}\right)\left(x_3^{f_3}\right)\ldots\left(x_n^{f_n}\right)} \quad (3)$$

or

$$\log(G) = \frac{\sum f_i \log(x_i)}{\sum f_i} \quad (4)$$

Where, G: Geometric Mean

x_n: weights of each respondents' number.

f_i: respondents number for each weights.

n: $\sum f_i$: the all respondent sum.

Table 1 Saaty Rating scale

Importance	Definition	Explanation
1	Equal importance of both elements	Two elements contribute equally
3	Moderate importance of one element over another	Experience and judgment favour one element over another
5	Strong importance of one element over another	An element is strongly favoured
7	Very strong importance of one element over another	An element is very strongly dominant
9	Extreme importance of one element over another	An element is favoured by at least an order of magnitude
2,4,6,8	Intermediate values	Used to compromise between two judgments

Table 2 Average random index according to matrix size [11]

Matrix Size(n)	1	2	3	4	5	6	7	8	9	10
Random Index	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

2.3 Analysis Grey Relational (AGR) [13]

(GRA) is a measurement method for discussing the consistency of discrete uncertain sequence and its target. It initiates by the original sequence factor creation, the measurement space factor using grey-relational generation, that is classified into three types:

- Smaller-the-better grey-relational generation: the minimum of the sequence factors is the ideal factor.
- Larger-the-better grey-relational generation: the maximum of the sequence factors is the ideal factor is.
- Nominal-the-better grey-relational generation: the one with the sequence factors target value in line is The ideal factor.

In grey relational generation, (5) used for normalized data corresponding to criterion, when Lower-the-Better (LB):

$$X_{i(k)}=\frac{\max y_{i\ (k)}-y_{i(k)}}{\max y_{i(k)}-\min y_{i(k)}} \quad (5)$$

Equation (6) used for the normalized data when Higher-the-Better (HB) criterion:

$$X_{i(k)}=\frac{y_{i(k)}-\min y_i}{\max y_{i(k)}-\min y_{i(k)}} \quad (6)$$

Where:

$X_{i(k)}$: is the value of $y_{i(k)}$ after the grey relational generation,

min $y_{i(k)}$: is the smallest value of $y_{i(k)}$, and max $y_{i(k)}$: is the largest value of $y_{i(k)}$.

The purpose of Grey relational grade is to appear the relation degrees between the sequences say, [$X_{0\,(k)}$ and $X_{i\,(k)}$, i = 1, 2, 3…, n].

The purpose of using GRA, in this research, is to normalize the values of realistic indicators by using (6) as the research intends maximum sustainability (higher is better).

3. Developing Economic and Social Sustainability Index

Reference [6] suggested steps for building the framework of composite indicators (index) as described in Figure 2. the steps for developing economic and social sustainability index in current study, are the following:

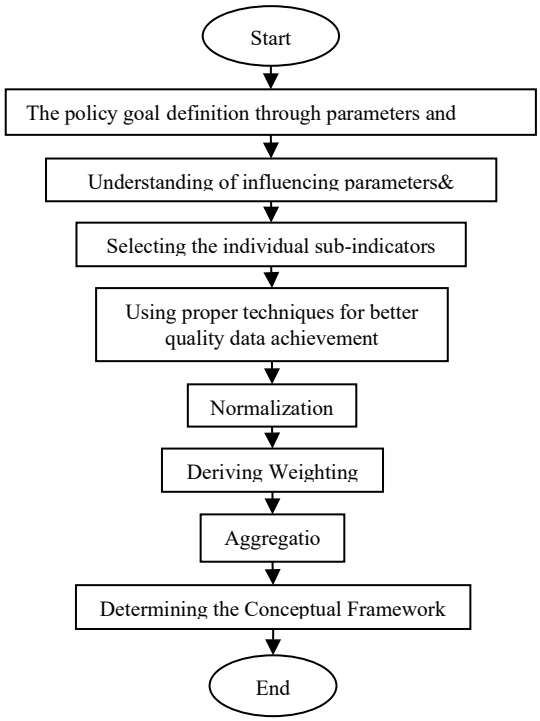


Figure 2: The Steps for establishing the framework of composite indicators [Researcher depending on [6]]

3.1. Developing Appropriate Sustainable Urban Framework

Reference [14] illustrated Developing three levels hierarchy Sustainable Urban framework for Iraqi cities (Dimensions, categories& indicators), as in Figure 3. This study deals with the economic and social dimension. Each dimension consists of two levels (categories level and indicators level). Six categories and eighteen indicators in The Economic dimension, and eight categories and forty-one indicators in The social dimension. the components for each dimension are shown in Table 3.

3.2. Determining Weightings Using AHP Method

Figure 4 illustrate the steps followed in the current research for determining framework components weighting by AHP method (Analysis Hierarchy Process).

The questioners used for finding the weight of categories, and indicators for economic and social dimension. The process of preparing questionnaire depends on research hierarchy framework and pair wise comparison. The questionnaire contains multiple issues and the answer accuracy needs a logical respondent with the following characteristics:

- Knowing about sustainability or urban sustainability.
- Involved in and active with projects and development of the governorate.

The researcher suggested to distribute at least (9) questionnaires for each governorate as in [15]. A list of the experts' respondents and their other information is summarized in Table 4. The number of received questionnaire respondents varies from one province to another, and the highest number of questionnaire received is (9) as in governorate of Baghdad and Salah Al-Deen and the lowest number is (4) as in the province of Wasit.

After accomplishing categories and indicators pairwise comparisons by the specialists' responses about the questionnaire, the aggregation for the individual judgments using the geometric mean as Saaty suggested in [16].

3.3. Determining Indicators Realistic Values

The researcher spent more than four months for collecting indicators realistic values for the study provinces. additional four months were needed for determining indicators values because of the difficulties to get fit value of indicator. The process of information gathering mainly depended on the Ministry of Planning / Central Statistical Organization & Information Technology (COSIT) in all its divisions, otherwise the values of indicators were obtained by other relevant ministries, authorities and departments, if information was unavailable. the process of determining values of indicators is as follows:

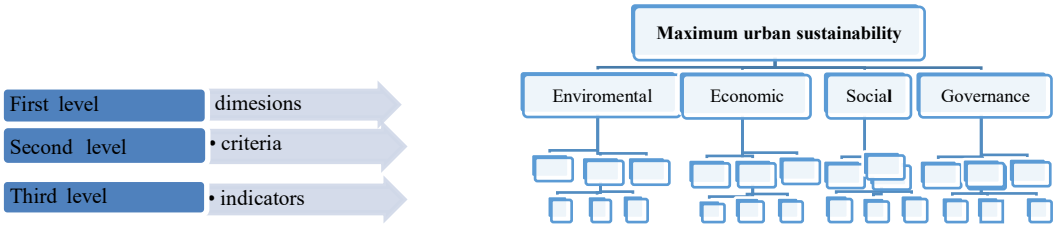


Figure 3: The hierarchy three levels Framework[Researcher]

Table 3: Economic and social sustainable framework component

Code	Details
2.	The economic dimension
2.1	The use of resources and energy
2.1.1	Electricity used per capita
2.1.2	Gasoline consumed per capita
2.1.3	Energy use per capita per year
2.2	Efficiency of economic development
2.2.1	Contribution of tourism to GDP
2.2.2	Investment share of GDP
2.2.3	Number of industrial projects
2.2.4	The share of government investment
2.3	Local government finance
2.3.1	Capital spending as a percent- age of total expenditures
2.3.2	Spending on research and development as a percentage of
2.3.3	Local government revenue
2.4	Services
2.4.1	Cost recovery
2.4.2	Loss in revenue as a result of power outages
2.5	Disasters
2.5.1	Percent of population living in areas at risk
2.5.2	Economic loss resulting from human and natural disasters
2.5.3	Safe design of the building of the topographic and geological
2.8	Economy population
2.8.1	Employment
2.8.2	Average family income
2.8.3	Per capita GDP
3.	The social dimension:
3.2	Health
3.2.1	The proportion of the number of hospitals and health
3.2.2	The proportion of the provision of potable water in urban areas, according to national standards
3.2.3	Hospital beds for every ten thousand people
3.2.4	rate of Mortality and birth
3.2.5	Number of doctors per capita
3.2.6	Access to sanitation facilities
3.2.7	Nutritional status of children
3.2.8	Immunization of children against disease infections
3.2.9	Prevalence of tobacco use
3.3	Education
3.3.1	The number of schools to population ratio
3.3.2	The number of researchers
3.3.3	Illiteracy rate for adults
3.3.4	Enrollment rates
3.3.5	The number of higher education graduates
3.3.6	The percentage of students who reach the fifth grade of primary education
3.3.7	The proportion of the number of students to teachers
3.3.8	The number of people who have graduated from junior
3.4	Housing
3.4.1	Housing shortage
3.4.2	The dimensions of the deficit housing and shortage ;neighborhoods, slums
3.4.3	Rent-to-income ratio
3.4.4	Optional decent housing or authorized
3.4.5	Type of tenure
3.4.6	Random housing
3.4.7	Abandoned housing
3.5	Social development and poverty eradication
3.5.1	percent of population living below the poverty line
3.5.2	The proportion of poor women , children with special needs who do not have access to community facilities and services
3.5.3	Child labor
3.5.4	Death under the age of five suffer from malnutrition
3.5.5	The proportion low-birth weight infants
3.5.6	The incidence of serious diseases
3.6	Violence and Crime
3.6.1	Violence in urban areas
3.6.2	Number of recorded crimes per 100,000 population
3.7	Urbanization and urban housing
3.7.1	The population growth rate
3.7.2	The number of families in one housing unit
3.7.3	Total net migration
3.7.4	The average household size for each family
3.7.5	The proportion of population living in urban areas
3.8	Culture of the community in sustainability
3.8.1	Public awareness of the population to hazardous
3.8.2	Percentage of the population who have gardens
3.9	Household Spending
3.9.1	Household spending
3.9.2	Average savings per household

- Some of the values were found directly.
- The researcher calculated some of these values with many equations and methods.
- Some indicators were calculated from some surveys that had a relationship to the indicators.
- Some indicators were new or that the relevant ministry responsible for providing the indicators values refused to provide the researcher with, so the researcher used an open interview questionnaire with Likert scale (i.e.,1: very low, 2: low, 3: medium ,4: high ,5: very high) to find its values.
- For the same reasons above, some of the economic indicators were found without monetary units.
- Some indicators included urban and rural regions together and some included urban only. In the first case, the indicator to be found was only for the urban through multiplication by the percentage of relevant urban in the governorate.

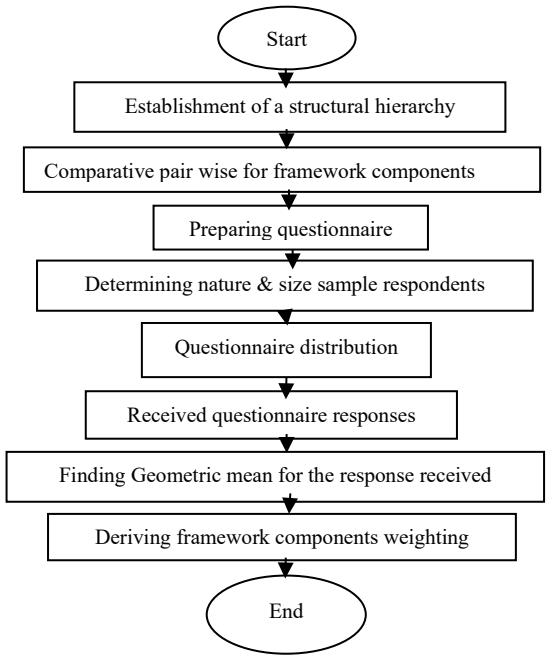


Figure 4: Flowchart for weighting framework components by AHP method [Researcher]

Table 4: Summary the respondents' information

provinces		Diala	Al-Anbar	Baghdad	Babylon	Kerbela	Wasit	Salah ALDeen
No. of respondents		8	5	9	7	5	4	9
Certificate	B.A.	2	0	0	0	2	2	3
	M.A.	2	0	4	3	2	0	1
	Ph.D.	4	5	5	4	1	2	5
profession	Planner	1	1	1	0	1	1	3
	Engineer	4	1	1	1	2	2	1
	Lecturer	4	3	3	6	2	1	5
Years of practice	10 to20	3	2	2	4	3	3	1
	more than 20	5	3	3	3	2	1	8

3.4. Indicators Influence identifying on Sustainability

Table 5 illustrated the direction of indicators against sustainability.

Table 5: Signs of indicators against sustainability [Researcher]

Signs of indicators	Means of the signs
-	If the indictor value increased sustainability decreased.
+	If the indictor value increased sustainability increased .

3.5. Using Analysis Grey Relational for Normalization indicators values

Because of the indicators values were not the same each other units, the normalization was performed to obtain one unit for all values of indicators, additionally to obtain sequential in the indicators values.

Equation (6) was used for normalization indicators values, the limits of indicators values normalization are between (0-1).

3.6. Determining Dimension Sustainability Index (DSI)by Aggregation

The base for determining the economic and social sustainability Indices is the aggregation as in (7).

$$DSI=\Sigma[W_c\ (\Sigma W_i X_i A_i)] \quad (7)$$

Where:
DSI: Dimension Sustainable Index
W_c: Weight of Category c; W_i: Weight of Indicator i;
X_i: Normalized obtained of indicator value i in the study area;
A_i: 1 or -1, that depends on indicator i influence on sustainable; i: indicator Number for the category c;
c: categories Number.

4. Results

4.1. Categories and Indicators weight for Economic and Social Dimensions

Tables 6 and 7 show the weights and ranks of economic dimension categories.

It is clear from the table that the category(C2.1) ranking in two provinces is the second and fourth ranking and the highest weight in the Babylon province. the category(C2.2) ranking in two provinces is the first ranking and the highest weight in the Wasit province. The category(C2.3) ranking in three provinces is the first ranking and the highest weight in the Salah ALDeen province and sixth rank in two other provinces. The category(C2.4) ranking in three provinces is the first ranking and the highest weight in the Diala province .and second rank in two other provinces. The category(C2.5) ranking in three provinces is the sixth ranking and the highest weight in the Salah ALDeen province .and second rank in two other provinces. The category(C2.8) ranking in three provinces is the fifth ranking and the highest weight in the Diala province, third and fourth rank in two other provinces. Tables 8,9 and10 show the weights and ranks of social dimension categories.

It is clear from the table that the category(C3.2) ranking in three provinces is the first and the highest weight in the Kerbela province. the category(C3.3) ranking in two provinces is the first ranking and the highest weight in the Wasit province and second rank in four other provinces. the category(C3.4) ranking in two provinces is the fourth ranking and the highest weight in the Al-Anbar province and fifth rank in two other provinces.

the category(C3.5) ranking in two provinces is the third ranking and the highest weight in the Diala province and fifth rank in three other provinces. the category(C3.6) ranking in two provinces is the sixth ranking and the highest weight in the Salah ALDeen province and seventh rank in three other provinces. the category(C3.7) ranking in three provinces is the eighth ranking and the highest weight in the Al-Anbar province. the category(C3.8) ranking in two provinces is the third ranking and the highest weight in the Diala province and fifth rank in two other provinces. the category(C3.9) ranking in two provinces is the seventh ranking and the highest weight in the Baghdad province and eighth rank in two other provinces. Table 11shows the priority summary for economic and social dimension categories.

Table 6: Weights and ranks of economic dimension categories

Category	C2.1		C2.2		C2.3	
Provinces	Rank	Value	Rank	Value	Rank	Value
Diala	2	0.157	5	0.115	4	0.117
Al-Anbar	1	0.072	2	0.078	1	0.283
Baghdad	4	0.109	6	0.055	1	0.179
Babylon	2	0.177	1	0.199	6	0.030
Kerbela	4	0.148	3	0.159	6	0.074
Wasit	3	0.153	1	0.241	2	0.213
Salah ALDeen	6	0.072	5	0.075	1	0.238

Table 7: Weights and ranks of economic dimension categories

Category	C2.4		C2.5		C2.8	
Provinces	Rank	Value	Rank	Value	Rank	Value
Diala	1	0.181	6	0.080	3	0.132
Al-Anbar	2	0.273	6	0.065	3	0.086
Baghdad	2	0.178	3	0.122	5	0.093
Babylon	5	0.075	3	0.144	4	0.142
Kerbela	1	0.163	2	0.160	5	0.111
Wasit	1	0.163	6	0.037	5	0.053
Salah Aldeen	3	0.174	2	0.200	4	0.098

Table 8: Weights and ranks of social dimension categories

Category	3.2		3.3		C3.4	
Provinces	Rank	Value	Rank	Value	Rank	Value
Diala	4	0.115	2	0.148	6	0.107
Al-Anbar	8	0.068	4	0.111	2	0.169
Baghdad	1	0.160	2	0.149	5	0.135
Babylon	1	0.261	2	0.176	5	0.089
Kerbela	1	0.206	2	0.133	3	0.121
Wasit	3	0.132	1	0.220	4	0.128
Salah Aldeen	7	0.093	1	0.163	4	0.121

4.2. Economic and Social Sustainability Index

In the current research, many mathematical equations, Microsoft Excel program was used. Table 12 shows a sample of sustainability index calculations for the economic dimension of Diyala province in excel sheet: After making calculations, Table13 illustrate the values of each sustainability dimension index in the current research. Figures 5and6 illustrate economic and social index for each province in study area. In the economic dimension, Baghdad got the first ranking followed by Wasit, and Salah ALDeen got the last ranking followed by Babylon. As for the social dimension, Baghdad got the first rank followed by Babylon and Wasit got the final ranking followed by Diyala.

Table 9: Weights and ranks of social dimension categories

Category	C3.5		C3.6		C3.7	
Provinces	Rank	Value	Rank	Value	Rank	Value
Diala	1	0.153	5	0.109	8	0.077
Al-Anbar	5	0.094	6	0.083	1	0.200
Baghdad	3	0.143	7	0.066	8	0.037
Babylon	5	0.089	7	0.068	3	0.167
Kerbela	5	0.109	6	0.102	4	0.111
Wasit	2	0.139	7	0.060	6	0.074
Salah Aldeen	3	0.128	2	0.151	8	0.062

Table 10: Weights and ranks of social dimension categories

Category	C3.8		C3.9	
Provinces	Rank	Value	Rank	Value
Diala	3.000	0.131	7	0.092
Al-Anbar	3.000	0.118	7	0.075
Baghdad	6.000	0.081	4	0.137
Babylon	4.000	0.105	8	0.036
Kerbela	7.000	0.097	8	0.076
Wasit	5.000	0.095	8	0.051
Salah ALDeen	5.000	0.113	6	0.113

Table 11: The priority summary for economic and social dimension categories.

Economic Category priority		Social Category priority	
Category	Rank	Category	Rank
C2.4	1	C3.2	1
C2.3	2	C3.3	2
C2.2	3	C3.5	3
C2.1	4	C3.8	4
C2.8	5	C3.4	5
C2.5	6	C3.6	6
		C3.7	7
		C3.9	8

2. Conclusions

In general, Degree of sustainability dimensions of Iraqi provinces, was very low: and Baghdad province got first in the economic and social index maybe it's because it is the capital. Salah ALDeen province got the last position in the economic index. Wasit province got the last position in the social index. The negative sign of the dimension index value means the sustainability was declined in that dimension. The economic index consisted of only one province with negative signal (Kerbela).

Table 12: A sample of economic sustainability index calculations of Diala

Code	Wc	Ci	Wi	Xi	Ai	WiXiAi	CI
2.1	0.157						
		2.1.1	0.416	0.026	-1	-0.011	
		2.1.2	0.200	0.001	-1	0.000	
		2.1.3	0.384	0.016	-1	-0.006	
					∑	-0.017	-0.003
2.2	0.115						
		2.2.1	0.228	0.000	1	0.000	
		2.2.2	0.211	0.049	1	0.010	
		2.2.3	0.377	0.028	1	0.011	
2.3	0.117					0.021	0.002
		2.2.4	0.185	0.023	1	0.004	
		2.3.1	0.295	0.051	1	0.015	
		2.3.2	0.431	0.013	1	0.006	
		2.3.3	0.274	0.037	1	0.010	
2.4	0.181					0.035	0.004
		2.4.1	0.439	0.016	1	0.007	
		2.4.2	0.561	0.006	-1	-0.003	
2.5	0.080					0.003	0.001
		2.5.1	0.209	1.000	-1	-0.209	
		2.5.2	0.284	0.600	-1	-0.171	
		2.5.3	0.506	0.506	1	0.256	
2.8	0.132					-0.124	-0.010
		2.8.1	0.184	0.084	1	0.015	
		2.8.2	0.494	0.282	1	0.139	
		2.8.3	0.323	0.065	1	0.021	
						0.176	0.023
						DI	**0.018**

Table 13: The values of economic and social index

Provinces	EconomicIndexDI2	Social IndexDI3
Diala	0.018	-0.116
Al-Anbar	0.02	-0.062
Baghdad	0.202	0.092
Babylon	0.003	0.084
Kerbela	-0.008	0.005
Wasit	0.046	-0.128
Salah ALDeen	-0.014	-0.044

The social index consisted of four provinces with negative values (Wasit, Diyala. Al-Anbar and Salah ALDeen). The dimension sustainability indexes (DSI) were varied from one province to another because of the index value depends on the categories and indicators weights and the indicators realistic value where the weights are based on the nature of region and the opinions of experts while the indicators realistic value depend on several things, as the following:

- The difference in topography and geography of the region, which contributes to different capacities.

- The difference in population density in each province, which affects the distribution of economic projects and development plans.
- The direction of the longitudinal transport is determined along the Tigris and Euphrates rivers, neglecting the horizontal transport, which affects the distribution of economic projects and development plans in the study provinces in addition to the other aspects (environmental and social).
- The priorities of the governorate determined by the central and local governments in the province.

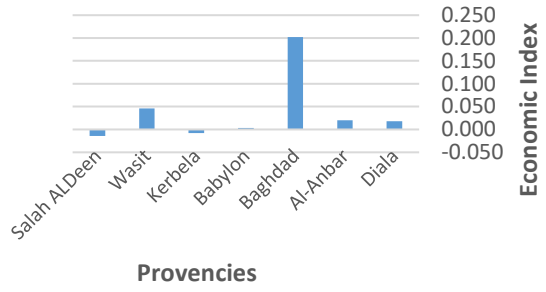


Figure 5: Economic indexes for each province in study area

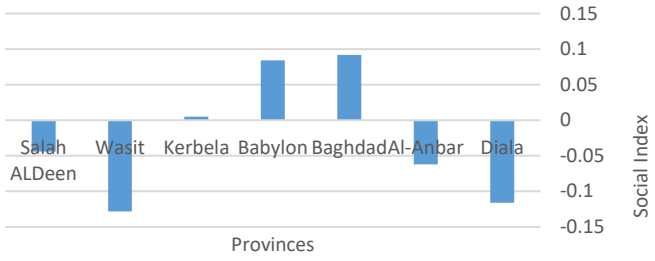


Figure 6: Social indexes for each province in study area

Conflict of Interest

There is many conflict such as respondents sample characteristics as mentioned in Section (3.2) are rare and if we find them, they will be very busy, so the researcher did the following steps for questionnaire distribution:

- Traveling for each governorate,
- Selecting most of the respondents,
- Distributing at least (9) questionnaires for each governorate, and
- Attending an open interview with respondents to ensure the understanding of questionnaire and to insuring the consistency of answers

Acknowledgment

I would like to thank Prof., Dr. Angham Ezzulddin Ali in the Department of Civil Engineering / University of Baghdad, Prof Dr. Mohammed Al-Quraishai and Dr. Abdul Wahab Ahmed at Urban and Regional planning centre /University of Bahadad, Prof. Dr. Numan Hussain and Prof. Dr.Sabah Faihan at Tikrit University ,Prof. Dr. Mohammed Yuosif at Diala university ,Prof.Adel Nuhair,Asst. Prof. Dr.Thair Shakeer at A-Anbar University , Dr.Azhar Hussain ,Dr.Mohamed Al-Said ,Dr Maher, Miss Huda Hadawi, Miss Ban and all the head of the departments at the Central Statistical Organization and Information Technology at the Ministry of Planning, and other Professors for their support and scientific assistance throughout this study. Great thanks to all persons including the esteemed professors, officials and engineers who gave me all the help and support in the universities and governmental institutions in the governorates.

References

[1] Portland State University, Twelfth International Conference on Environmental Cultural, Economic & Social Sustainability "Urban Sustainability – Inspiration and Solution", Portland State University | Portland, USA, 2016.

[2] Environment Management Group Wang, A Framework For Advancing Environmental And Social Sustainability In The United Nations System, United Nations, Geneva, 2012.

[3] Spangenberg, J.H). "The changing contribution of unpaid work to the total standard of living in sustainable development scenarios". Int. J. Sustainable Development, 5(4);461–475,2002.

[4] سليمان مهنا, ريدة ديب, التخطيط من اجل الاستدامة, مجلة جامعة دمشق للعلوم الهندسية, المجلد 25 العدد 1, 2009

[5] Hsiang-Hsuan Chih, Yu-En Lin & Pin-Huang Chou, "Disposition Effect, Escalation of Commitment and Mutual Fund Performance".Management Review methodology,28 (4); 103-107, 2009.

[6] Tanguay, G. Rajaonson, U., Lefebvre, J.F., Lanoie, P.," Measuring the Sustainability of Cities: A Survey-Based Analysis of the Use of Local Indicators". Ecological Indicators;10: 407-418,2009. Available on www.cirano.qc.ca /pdf/publication/2009s-02.pdf .

[7] Singh, R.K., Murty, H.R., Gupta, S.K., Dikshit, A.K., "An overview of sustainability assessment methodologies". Ecological Indicators , 15; 281–299,2012.DOI of original article:10.1016/j.ecolind.2008.05.011. E-mail address: r.singh@pe- international.com (R.K. Singh). (Meadows, 1998). 1470-160X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi: 10.1016/j.ecolind.2011.01.007.

[8] Loken, E. . Use of multi-criteria decision analysis methods for energy planning problems. Renewable and Sustainable Energy Reviews, 11(7): 1584-1595,2007.

[9] Kendrick, J. D. &Saaty, D., "Use Analytical Hierarchy Process For Project Selection". ASQ Six Sigma Forum Magazine, 6(4); 22-29 ,2007.

[10] Saaty, T.L.. A scaling method for priorities in a hierarchichal structure. Journal of Mathematical Psychology, 15, 234–281,1977.

[11] Saaty, T.L. . Fundamentals of the Analytic Hierarchy Process, RWS Publications, Pittsburgh, PA,2000.

[12] محمد أماني موسى, "التحليل الإحصائي للبيانات". مركز تطوير الدراسات العليا والبحوث، كلية الهندسة، جامعة القاهرة، الطبعة الأولى 2007. Available in www.capscu.com.

[13] Lin, S.T., Horng, S.J., Lee, B.H., Fan, P., Pan,Y., Lai, J.L, Chen, R.J and Khan,M.K.."Application of Grey-relational Analysis to Find the Most Suitable Watermarking Scheme", International Journal of Innovative Computing, Information and Control,7(9): 5389-5401,2011. ICIC International 2011 ISSN 1349-4198

[14] Mansor, M. A.& Mohammed ,S. R. "Developing Sustainable Urban Planning Indicators Framework for Iraqi Cities by Using Content Analysis and Delphi Method" International Journal of Modern Sciences and Engineering Technology (IJMSET)ISSN 2349-3755; Volume 2, Issue 3, pp.46-56, 2015.

[15] Akadiri,O.P., "Development of a Multi-criteria Approach for the Selection of Sustainable Materials for Building Projects". Ph.D. thesis, School of Engineering and the Built Environment, University of Wolverhampton, UK ,2011.

[16] Saaty, T.L., "Decision Making for Leaders: The Analytic Hierarchy Process for Decisions in a Complex World". RWS Publications, Pittsburgh, 2001.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 454-458 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Minimally Invasive, Thermal Energy Based, Cost-Efficient Method to Measure Fluid Flows in Compact Systems

Saurin Patel*, Rick Walker

SMC Ltd., Product Design and Development, 44087, USA

ARTICLE INFO

Article history:
Received: 28 August, 2018
Accepted: 01 October, 2018
Online: 25 October, 2018

Keywords:
Flow meter
Thermal Fluid Sensor
Non-intrusive
Minimally intrusive
Heat Transfer Coefficient
Cooling Rate
Affordable

ABSTRACT

Flow measurements are important for numerous industries as it directly relates to proper equipment functioning and product quality. This work demonstrates a novel thermal energy based compact flow measuring sensor. The sensor injects thermal energy into the core which is then extracted by the moving flow through the center of the core. Temperature changes produced by thermal energy exchange is evaluated over time and then used to calibrate and measure flows. The sensor can detect extremely low energy changes therefore measure relatively low flow rates as well. Being minimally intrusive, it does not disturb the flow eliminating any additional resistance commonly introduced by intrusive type sensors. It is made from inexpensive components and is easy to assemble therefore making it affordable. The simple geometry allows for it to be easily incorporated into compact systems and can be used in miniature devices to measure low flows. We declare this minimally intrusive only due to breaking the flow path with the sensor element otherwise it is non-intrusive.

1. Introduction

Fluid flow measurement is an essential as well as critical step in numerous processes and finds applications in many fields such as medical diagnostic devices, pharmaceutical industry, food industry, automotive industry, chemical industry, compressed air systems, and others (1). Precise fluid management is essential to regulate basic functionalities of a system. In certain applications accurate flow measurement becomes the key player by directly influencing the quality of the product (2). Rapid research and developments in microfluidics also dictates a need to develop ways to accurately measure and analyze extremely low flow rates (3). Flow measuring can be classified broadly into two categories: intrusive and non or minimally intrusive. In an intrusive method, the flow encounters a constriction or other mechanical interference. Intrusive methods typically depend on the forces produced by the flow which are then correlated to flow rate values. Mechanical flow meters, pressure dependent meters, and vortex based flow meters are examples of intrusive type flow meters. These meter types are relatively inexpensive and are available in variety of shapes and sizes. However, they are limited by their repeatability, accuracy and operating ranges. Being intrusive they reduce the flow rate within the system due to added resistance above the inherent system flow resistance. At low flow rates the kinetic energy associated is also low and inadequate to be converted into measureable mechanical movement (4). Fluids temperature changes affect the internal structure of these types of meters rendering this type of devices not suitable for applications where fluid temperatures changes are frequent and occupy a wide range (5).

Non-intrusive methods include optical, sonar, thermal, electromagnetic, Coriolis, and laser Doppler based flow meters. These types of flow meters do not interfere with the fluid motion or path and can measure most liquids. However electromagnetic flow meters can only measure the flow rate of a fluid conducting electricity and are expensive compared to their intrusive counterparts (6). Other flow meters based on optical, sonar, and Coriolis methods are accurate but very expensive to implement and have limited capabilities at low flow rate (6). Traditional thermal sensors need complex control system involving pulse width modulation (PWM) to constantly provide power to the heated element and maintain its temperature (6). The energy loss from the heated element due to fluid flow is correlated to the flow rate for measurements and requires extensive data acquisition. We herein present a novel way to measure flow rate and fluid velocity in a minimally intrusive, easy to assemble and cost-effective manner by simple cooling rate measurement obtained from temperature readings. The sensor only needs to be powered once for a reading, measures the energy loss from the system and establishes correlation with the flow rate. Moreover, the sensor was very easy to calibrate, capable of detecting low flows and has excellent

*Saurin H. Patel, Twinsburg Ohio, +1 (330) 342 7800, saurin.patel@smcltd.com

linearity and response time. In addition to the advantages mentioned above, the sensor is calibrated at a defined inlet temperature and therefore accounts for fluid property variations resulting in accurate readings.

2. Experimental Method

2.1. Material Selection and Sensor Design

A cylindrical aluminum rod with dimensions as shown in Figure 1a and 1b was used as the central core for the sensor. The density of aluminum being low compared to other conductive metals is capable of producing rapid temperature changes with small amount of stored thermal energy loss thus enabling low flow rates to produce measurable temperature differences. A circular opening equivalent to the inner diameter of the polymer tube was machined first at the center of the block for the fluid pathway. Counter boring operation was performed on the two flat outer surfaces for a short length to create an opening equal to the outer diameter of the polymer tubing. The tubing was then inserted from both sides and is represented in Figure 1a. The resulting geometry ensured uniform opening at the center of the aluminum core for the fluid motion, with minimal flow disturbances and alteration to system flow rate. Figure 1b shows the side view of the aluminum core. An opening with its center at the midway of inner diameter and outer diameter of the block was machined to accommodate the thermal sensor. Before inserting the sensor, adequate thermal grease was introduced into the opening to maintain thermal contact between the aluminum core and the sensor. The *Biot number* associated with the selected block is small (<0.1) which signifies analysis based on lumped heat capacitance method for the current system is valid with an error of less than 5%. Therefore, it is safe to assume that the temperature measured at any location on the aluminum block is at least 95% accurate and closely resembles the temperature of the entire block.

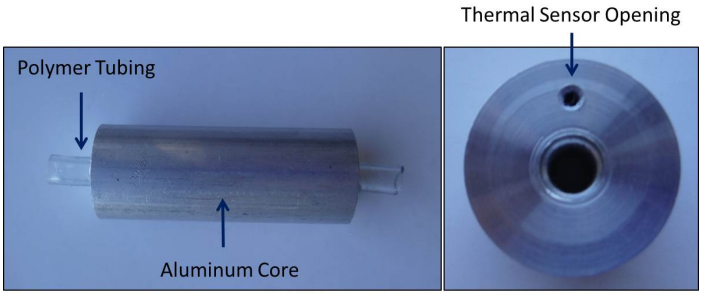


Figure 1: (a) Aluminum core with connected polymer tubes and (b) side view with counter bore and thermal sensor opening

2.2. Data Acquisition

Figure 2a represents the resistive heating element utilized to supply thermal energy to the aluminum core. A 12V DC power supply (YH-302D Yihua, China) connected to the heating element was controlled by Arduino UNO that used the temperature readings obtained from the thermal sensor (10K thermistor, see Figure 2b). The thermal sensor was calibrated using the Steinhart-Hart equation which is capable of producing accurate results between the ranges of 0 - 50° Celsius (Error -0.005 – 0.002° C). Code was written and embedded into to the Arduino UNO to monitor the temperature readings continuously and perform logical and arithmetic operations as needed. Initially, fluid was cooled to a desired temperature (14°C and 15°C in the present testing) using a novel system which also contained a pump to produce necessary volumetric flow rate. The system continuously monitored fluid temperature just before it entered into the sensors' aluminum core. Once fluid reached desired stable temperature a constant flow was allowed to pass through the core and the flow rate was measured using a calibrated flow meter (intrusive). The embedded code then turned the power supply on when the temperature of the aluminum core dropped to 16° C and turned the power supply off when the temperature reached 21° C. The constant flow of cold fluid dropped the temperature of the core. Time required by the fluid moving at a particular flow rate (recorded from the intrusive flow meter) to drop the temperature of the aluminum core by 2° C from 20 to 18° C was recorded. Cooling rate of the flow was calculated using (8). The flow rate to the aluminum sensor core was then randomly altered by pinching the polymer tube and all the previously mentioned steps were repeated for four different flow rates and two different inlet flow temperatures. The cooling rates obtained from the code and volumetric flow rates from the flow sensor were tabulated and are presented in the results and discussion section. The correlation obtained was then utilized to calibrate the minimally intrusive sensor.

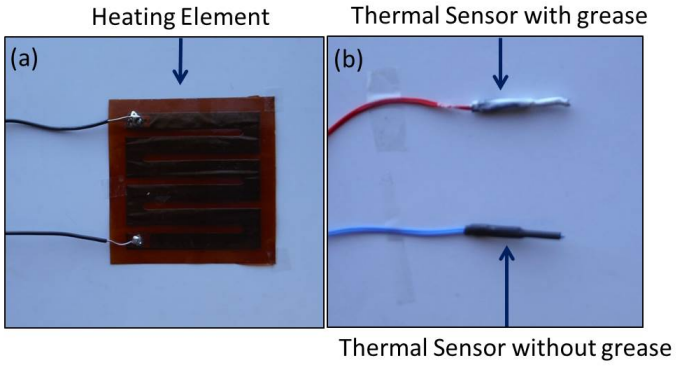


Figure 2: (a) Heating element with cables that are to be connected to a DC power supply and (b) thermal sensor used to monitor temperature readings (with and without thermal grease)

3. Theory

3.1. Thermal Analysis

The heater element attached to the peripheral surface of the aluminum core elevates the temperature of the entire metal by converting electrical energy into thermal energy. Thermal energy received by the aluminum core from the outer surface is conducted inwards towards the center. The energy reaching the center is carried away by the fluid making contact with the aluminum core through the mode of convective heat transfer as represented in Figure 3.

$$\dot{E}_{st} = \dot{E}_{in} - \dot{E}_{out} \quad (1)$$

$$\dot{E}_{st} = \dot{E}_{out} \quad (2)$$

The amount of energy stored is equal to vector summation of energy entering and exiting the system in the absence of internal energy generation. The energy balance of flow rate measuring system is represented by (1). When the temperature of the aluminum core as measured by the thermal sensor reaches 21° Celsius, the power supply to the heater is terminated. As a result the energy entering the system is effectively zero from that moment onwards. Eliminating the energy entering term (1) can be condensed into (2) which suggest that the energy stored in the aluminum core is gradually conveyed away by the energy exiting

the flow rate measuring system. In the present system, energy can exit through three different medium namely (a) convective heat transfer – water and (b) convective heat transfer – air and (c) radiative heat loss. The insulating layer of Poron© covering the entire outer aluminum core surface minimizes the convective heat

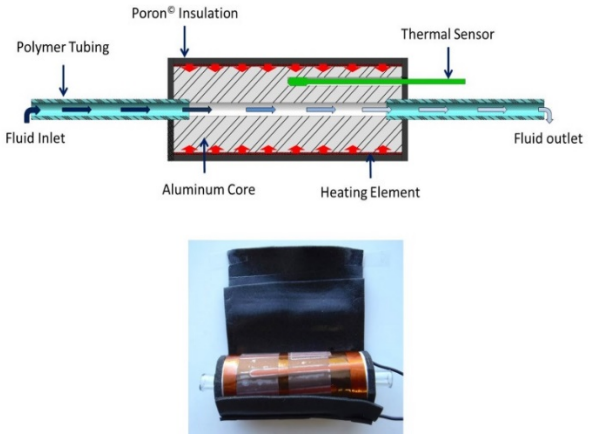


Figure 3: Schematic representation of the sensor including the heater element, aluminum core, thermal sensor, fluid tubing, and insulation. The sensor used for experimentation is also shown in the image.

transfer produced by air flow and therefore can be neglected. The radiative heat transfer losses for the temperatures at which the current system operates is also minimal and can be neglected as well. Eliminating other modes of heat transfer it is safe to conclude that the energy stored in the aluminum rod after the power supply to the heater is terminated is conveyed out by the fluid (14°C and 15°C) flowing through the center of the core. The energy balance between stored energy in the aluminum core and fluid motion is represented in (3). The amount of energy stored in the fixed mass aluminum core can be quantified as shown in (3) where m_{al}, C_P, ΔT_{al}, h, A, ΔT_f, and Δt are mass, heat capacity, temperature difference in the aluminum core, convective heat transfer coefficient, area fluid makes contact with the aluminum core, temperature difference between fluid entering and exiting the center and time interval respectively (7).

$$m_{al}C_p\left[\frac{\Delta T_{al}}{\Delta t}\right] = hA[\Delta T_f] \tag{3}$$

Mass, heat capacity, area of contact, and the temperature difference of the aluminum core (2° Celsius) in the current test system can be termed constant. Moreover, the system is designed to produce minimum variation in the fluid temperature exiting the core by utilizing a relatively small length of the core and thus temperature difference of fluid can be assumed constant for most practical purposes. Arranging all the constants on one side and variables on the other (3) can be represented in a different form as shown in (4). The modified equation dictates that the product of convective heat transfer coefficient and time interval is practically a constant. In other words, the fluid condition for which the convective heat transfer coefficient is large, requires a smaller time interval to extract stored energy compared to a fluid condition with smaller convective heat transfer coefficient for the above mentioned flow rate measuring system. The correlation suggests, with the increase in convective heat transfer coefficient of the fluid the time needed to cool the aluminum core by 2° Celsius will decrease.

$$\frac{m_{al}C_p\Delta T_{al}}{A\Delta T_f} = h\Delta t \tag{4}$$

The convective heat transfer coefficient based of Nusselt number evaluated using Gnielinski correlation that accounts for friction factor (surface roughness) f, Reynolds number Re_D, Prantl number Pr, thermal conductivity of fluid k, and hydraulic diameter D is given by (5). The correlation in (5) can be used with the hydraulic diameter, independent of the surface thermal condition (7). As seen from (5), the convective heat transfer coefficient is directly proportional to the Reynolds number of the fluid which in turn is proportional to the flow rate of the fluid (6). Q, ϑ and A_C are volumetric flow rate, kinematic viscosity, and cross-sectional area respectively in (6). Based on the above analysis we can deduce that with the increase in flow rate the convective heat transfer coefficient associated with the fluid also increases. The proposed minimally invasive flow rate measuring sensor exploits the above mentioned correlation between the fluid flow rate and the convective heat transfer coefficient.

$$h = \frac{\left(\frac{f}{8}\right)(Re_D - 1000)Pr}{1 + 12.7\left(\frac{f}{8}\right)^{0.5}(Pr^{0.67} - 1)}\frac{k}{D} \tag{5}$$

$$Re_D = \frac{QD}{\vartheta A_C} \tag{6}$$

$$h \propto Re_D \propto Q \tag{7}$$

3.2. Working Principle - Sensor

In order to quantify the energy exiting the aluminum core at different volumetric flow rate, a cooling rate term $\Delta T_{al}/\Delta t$ was defined that is a direct function of convective heat transfer coefficient h. The cooling rate as shown in (8) can be defined as the decrease in temperature of the aluminum core per unit time. From the above analysis we predict that with the increase in flow rate the convective heat transfer coefficient increases and therefore the cooling rate.

$$\left[\frac{\Delta T_{al}}{\Delta t}\right] = h\left[\frac{A\Delta T_f}{m_{al}C_p}\right] \tag{8}$$

Theoretically from (8) we can say that the cooling rate or energy loss from the aluminum core will increase with an increase in the convective heat transfer coefficient and the other terms on the right side of the equation are practically constant therefore acting as a multiplication factor only. Cooling rate at several different flow rates and couple different inlet fluid temperatures (14°C and 15°C) were experimentally obtained and tabulated. The results are discussed in the next section. The experimentally obtained data was then utilized as calibration factor to evaluate flow rate in a minimally invasive manner.

4. Results and Discussion

After Figure 4 represents the relationship between volumetric flow rates through the sensor core and the associated cooling rates obtained from the code for inlet fluid temperature of 14° Celsius. A multiplication factor was used to represent the cooling rates with a better resolution. Time measurements for the flow to reduce the aluminum core temperature by 2° Celsius were extracted from the code and manual calculations were performed to cross check the cooling rates obtained directly from the code. The cooling rates obtained using both the methods were found to be identical. As can be seen from the experimentally obtained plot, the cooling rate increases with the increase in volumetric flow rate. The

experimental results are consistent with the theoretical evaluation performed previously (7 & 8) which predicted that the cooling rate is proportional to the flow convective heat transfer coefficient h, which in turn is proportional to the volumetric flow rate Q. Additionally, it is important to note that the relationship between the cooling rate and volumetric flow rate was linear for the tested flow rates. We do acknowledge minor error in flow rate measurements originating from the accuracy of the intrusive flow meter. The error in the experimental measurements is accounted by the error bars as represented in the plots. Utilizing a Flow meter with greater accuracy can be used to obtain more accurate results for the calibration. Nevertheless, the result predicts a linear relation between the entities of concern in the given test performed.

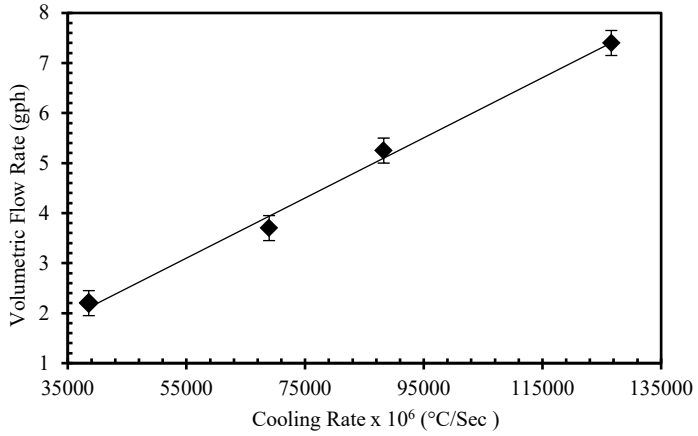


Figure 4: Cooling rates obtained from the code at various volumetric flow rates for inlet fluid temperature of 14° C

Figure 5 represents the relationship between volumetric flow rates through the sensor core and the associated cooling rates obtained from the code for inlet fluid temperature of 15° Celsius. Consistent with the earlier study for inlet temperature of 14° Celsius the cooling rate increases with the increase in volumetric flow rate through the aluminum core. Moreover, the coorelation between the entities for 15° Celsius was also linear. Suggesting cooling rate as defined earlier can be used as a good calibration factor to measure flow rates and velocity (from fluid properties and sensor geometry) in compact systems using a minimally intrusive method.

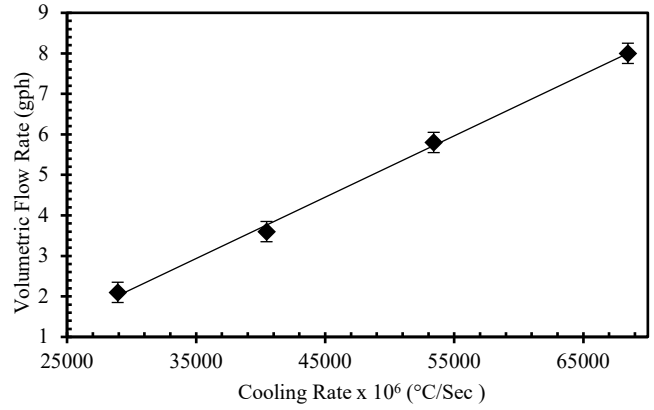


Figure 5: Cooling rates obtained from the code at various volumetric flow rates for inlet fluid temperature of 15° C

On comparison of the two plots presented, the cooling rates for the 14° Celsius inlet fluid temperature are higher than the 15° Celsius for a given volumetric flow rate. This can be expected as a colder fluid can extract more thermal energy from the aluminum core compared to a relatively warm fluid in a defined time interval. Therefore, the time required for a colder fluid to reduce the core temperature by 2° Celsius was smaller compared to a warmer fluid which results in a higher cooling rate (8). The correlation obtained for the two different fluid inlet temperatures were programmed back into the code and volumetric flow rates were measured using them. The relation produced accurate readings in both the cases.

The slope values calculated from the plots above for the 14° Celsius and 15° Celsius inlet fluid temperatures are 200 °C/sec/gph and 60°C/sec/gph respectively. If we compare the slopes of the two plots, it can be seen than the colder fluid (14° Celsius) has a smaller slope compared to the warmer fluid (15° Celsius) reconfirming the cooling rate was higher for a colder fluid compared with a warmer fluid. Analysis the slope valves, we can also observe that the slope values decrease with the decrease in the inlet fluid temperature. Knowing that the relation between the cooling rate and flow rate is linear for both the inlet fluid temperature and realizing that the cooling rate values at no flow condition will be same (i.e. small cooling rate at no flow condition defining the origin for the curves) for any inlet fluid temperature, we can define a single calibration factor or correlation by collecting adequate experimental data. The correlation can then be used to develop a single sensor that can identify volumetric flow rates for a broad range of inlet fluid temperature conditions.

5. Conclusion

We have developed a prototype sensor to measure volumetric flow rate and fluid velocity in compact flow systems using a minimally intrusive method for a defined inlet fluid temperature. The sensor was developed using an aluminum core that is heated and is allowed to be cooled by convective heat transfer produced by the fluid motion. The time associated with the aluminum core cooling was used to define the calibration factor and eventually flow rate. The linear relationship between the cooling rate and volumetric flow rate can be effectively exploited to develop a universal sensor that can measure flow rates for several different inlet fluid temperatures. The absence of any mechanical components or flow restrictors in the sensor's measurement path prevents any volumetric flow rate modifications within the system. No flow restriction in the system from the sensor means the fluid driving mechanism (e.g. pumps & motors) can be efficiently used as it does not have to overcome the additional losses created by the presence of sensor restriction. Moreover, the sensor uses commonly available items making it extremely economical to fabricate and easy to assemble providing a significant advantage over expensive ultra-sonic, electromagnetic, Coriolis and other flow meters. However, we appreciate the fact that the response time of the system can be slow for extremely low flow rates and the sensors need to be calibrated for the operating temperature.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is supported by SMC Ltd. Product Design and Development, Twinsburg, Ohio. We thank our colleagues from Cerêve, Inc. who provided their novel fluid temperature regulating device that greatly assisted the research, although they may not agree with all of the conclusions of this paper.

References

[1] D. N. Ku, "Blood flow in arteries" Annual Review in Fluid Mechanics. 21, 399–434, 1997. https://doi.org/10.1146/annurev.fluid.29.1.399

[2] L. X. Yu, "Pharmaceutical Qulaity by Design: Product and Process Development, Understanding, and Control" Pharma Res., 25, 2463-2463, No. 10, 2008, https://doi.org/10.1007/s11095-007-9511-1

[3] S. Patel, D. Showers, P. Vedantam, T-R. Tzeng, S. Qian, X. Xuan "Microfluidic separation of live and dead yeast cells using reservoir-based dielectrophoresis" Biomicrofluidics, 6(3), 034102, 2012, https://doi.org/10.1063/1.4732800

[4] J. G. Santiago, S. T. Wereley, C. D. Meinhart, D. J. Beebe, R. J. Adrian, "A particle image velocimetry system for microfluidics" Experiments in Fluids, 25, 316-319., 1998, https://doi.org/10.1007/s003480050235

[5] H.A. Stone, A.D. Stroock, A. Ajdari, Engineering flows in small devices: "Microfluidics towards a Lob-on-a-Chip" Annual Review in Fluid Mechanics. 36:1, 381-411, 2004, https://doi.org/10.1146/annurev.fluid.36.050802.122124

[6] R. C. Baker, Flow measurement handbook, Cambridge University Press, 2016

[7] T. L. Bergman, A. S. Lavine, F P. Incropera, D. P. DeWitt, Fundamentals of Heat and Mass Transfer, 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 455-468 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Simulation-Optimisation of a Granularity Controlled Consumer Supply Network Using Genetic Algorithms

Zeinab Hajiabolhasani [*,1,2], Romeo Marian[2], John Boland[1]

[1]*School of Information Technologies and Mathematical Sciences, University of South Australia, 5095, Australia*

[2]*School of Engineering, University of South Australia, 5095, Australia*

ARTICLE INFO

Article history:
Received: 29 August, 2018
Accepted: 05 December, 2018
Online: 20 December, 2018

Keywords:
Consumer Supply Network
Simulation-Optimisation
Granularity
Genetic Algorithms

ABSTRACT

The decision support systems regarding the Supply Chains (SCs) management services can be significantly improved if an effective viable method is utilised. This paper presents a robust simulation optimisation approach (SOA) for the design and analysis of a granularity controlled and complex system known as Consumer Supply Network (CSN) incorporating uncertain demand and capacity. Minimising the total cost of running the network, calculating optimum values of orders and optimum capacity of the inventory associated with each product family are the objectives pursued in this study. A mixed integer non-linear programming (MINLP) model was formulated, mathematically described, simulated and optimised using Genetic Algorithms (GA). Also, the influence of the problem's attributes (e.g. product classes, consumers, various planning horizons), and controllable parameters of the search algorithm (e.g. size of the population, crossover rate, and mutation rate) as well as the mutual interaction of various dependencies on the quality of the solution was scrutinised using Taguchi method along with regression. The robustness of the proposed SOA was demonstrated by a series of representative case studies.

1. Introduction

The main challenges affecting today's Supply Chains (SCs) are globalisation, environmental and technological turbulences and rapid changes in economy capacity. They have provoked companies to recognise that, in order to remain competitive in the global market, they need to gain more from their SCs.

Supply Chains are defined as links (relationships) between every unit (enterprise) in a manufacturing process from raw materials to customers. Traditionally, products were made and flowed to consumers through SCs. However, due to globalisation and complexity of the economy, today's SCs are better characterised as Supply Networks (SNs).

Consumer Supply Networks (CSNs) refer to complex networks consisting of sets of companies working in unison to supply, manufacture, distribute and deliver final products and services to end-users (Figure 1), being controlled by information flow.

CSNs are examples of industrial systems that are naturally large, complex, stochastic, and dynamic. These attributes translate into difficulties in representing the actual behaviour and in planning, optimising and anticipating performance. Also, the combination of these attributes makes the choice of an appropriate solution methodology difficult at best, if not simply impossible at this point in time [1].

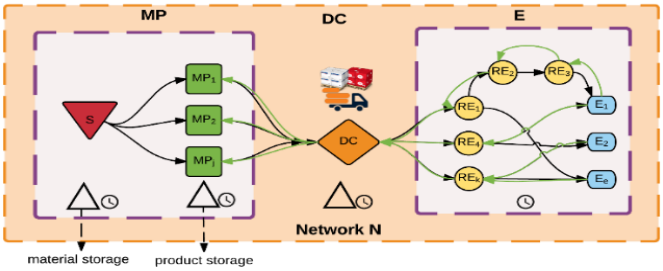


Figure 1 Three echelon Consumer Supply Network

Different methodologies have been utilised to solve this class of complex problem; simulation and optimisation methods are widely used to tackle such problems.

Simulation is a powerful tool for modelling, analysis, and validation of CSNs. However, its major disadvantage is that it will produce a very detailed analysis but strictly for a given

* Zeinab Hajiabolhasani, Email: zeinab.hajiabolhasani@unisa.edu.au

configuration. Simulation cannot change the configuration of the system, and any optimisation would be searching for the best combination of variables for a given system.

A recurrent, key issue when attempting to optimise CSN is the granularity of the model. An appropriate granularity – the size of the smallest indivisible unit (of product, part, flow, time, etc.) of the process – makes the difference between a successful implementation of the optimisation methodology and an algorithm that does not converge or gets consistently stuck in local optima. Additionally, the choice of the granularity of the model has to be easy to translate in practice – a purely theoretical solution that cannot be implemented in real life is of little help.

This paper is an extension of the work initially has been presented in Intellisys Conference [2] in which a unique simulation optimisation approach (SOA) within an integrated methodology was developed. A small-scale Multi-Period, Multi-Product Consumer Supply Network (MPMPCSN) model, using mixed integer non-linear programming (MINLP) was designed. Then, the optimum quantity of orders was determined incorporating GA optimisation algorithm which simultaneously results in the total inventory cost minimisation. This way, the unique advantages of simulation were incorporated with optimisation method and higher quality solutions were achieved. Also, the quality of the solutions that were obtained by the proposed framework was checked by fine-tuning of the search algorithm's parameters combining the simulation model with the Taguchi method. Hence, in this study, a series of computational trials on realistic test problems are designed and analysed to demonstrate the generalisability of the proposed SOA for problems of similar size at different granularity levels.

The rest of this article is organised as follows: Section 2 is devoted to reviewing modelling methodologies that were used to solve CSNs problems. Section 3 presents the proposed MINLP model. Section 4 provides details about granularity. The optimisation module of the SOA methodology is described in Section 4. The numerical examples are given in Section 6 and discussed in Section 7. Section 8 concludes the paper.

2. Literature Review

A number of potential solution methods for the class of problems of similar size and complexity have been developed in the literature ranging from classical mathematical programming to hybrid and systematic methods [1, 3].

Optimisation methodologies combined with mathematical models are mainly contributed to solutions validation. A stable optimal solution can be obtained by a given objective function subject to several constraints. However, they are unable to provide the gradient of design space over time [4]. The extent of the optimisation problem cannot be expanded beyond a certain limit as the complexity of the problem adversely affects the computational costs which make less efficient and less practical [5]. This concern can be addressed by using, simulation methodologies.

Simulation models can deal with all attributes of CSNs problems which makes them a powerful analytical tool in this area [6]. In particular, CSN simulation provides a model that suitably represents, processes associated with specific business units such as ordering system, manufacturing plant, distribution centres, etc. in the presence of uncertainty [7, 8]. Simulation modelling methods alongside with mathematical and models based on algorithms almost always come together. The main advantage of simulation approaches is a possibility to explore *what-if* scenarios that provide a deeper understanding of the dependencies in a system. The operations of a real system that are usually very dangerous, expensive, or impractical to implement can be evaluated according to their resilience and robustness subject to various predefined inputs (e.g. time horizon, resources, etc.) and at any desired granular level via simulation modelling. Using computer programming, the performance of a real system subject to controlled and environmental changes can be simulated. Therefore, many input values and their combinations can be explored through simulation models [9]. Also, simulation models offer flexibility in developing and assessment of different scenarios, with reasonably high-speed processing. In addition, an embedded standard reporting system make them unique in modelling, analysing, and validating of complex systems.

As pointed out, independent deployment of optimisation and simulation methodologies has some benefits. However, it also has limitations. The main drawback of simulation models is that they can only work with a set configuration of a solution. On the other hand, finding the optimal solution by independently using the traditional optimisation approaches incurs heavy computational cost. Therefore, the integration of the two methods may lead to a uniquely efficient optimisation.

SOA is a key factor of modern design across industries [3]. SOA is often used in the design, modelling and in analyses of systems. It can provide an optimal setting for set of parameters for a simulation model [10]. But due to high computational requirements, scientists have not given much attention to the use of SOA in CSNs [10-12]. Consequently, SOA turns into a hot research topic for optimisation of CSNs. The optimisation core together with a simulation model in SOA, can search the solution space globally (ergodicity of GA) whereas the simulation module acts as a quality assessment unit.

Following the advances in computational power, increased efforts have been made to leverage simulation for optimisation/simulation-based optimisation of hybrid systems with behaviours that can be discrete or continuous [13]. CSNs are hybrid systems with a high level of complexity.

Inventory control planning problems have been tackled using many metaheuristic algorithms [5, 14-18]. GA was widely used to solve related problems [19]. Through exploring the solution space, GA finds optimal or near optimal solutions. But, like in other evolutionary algorithms (EA), GA cannot carry out self-validation. GA risks to converge to local optima [20]. Hence, a valid question is whether or not the obtained solution is a high-quality candidate.

The parameters of the search algorithm - population size, crossover and mutation and rates, as well as the interaction between these parameters have significant impacts on the quality of solutions. As the entire search population or its fitness function might be highly affected by variation of these parameters. This necessitates implementation of a mechanism that can offer parameters tunning is essential. However, it is very hard to perform perfect tuning due to complexity among the interactions of EA's parameters. Most often, trial and error of EA's parameters is used in OR studies. However, experimentally tuning the parameters is less practical and very expensive [21]. We thus, propose using

statistical methods based on experiments as a more robust approach [22].

In [23], the authors present a multi-echelon SN simulation-based optimisation model for a multi-criteria P-D design. The model offers concurrent optimisation of the network's structure, the set of the control strategies, and the quantitative parameters of the strategy for control. The modelling, simulation and then optimisation of networked entities are performed using a graphical interface designed in C++ programming. In this study, the candidate solutions are evaluated by a discrete-event simulation (DES) module. A multi-objective GA algorithm is developed aiming at finding compromised solutions regarding structural, qualitative and quantitative variables. The toolbox developed in the research considers a real Production-Distribution model which makes it a unique decision support system. However, there is no evidence shown with regards to parameters tuning of the GA algorithm.

In [24], the authors describe a two-phase Mixed Integer Linear Programming model addressing planning and scheduling systems of a build-to-order SN system. They use GA to optimise the aggregate costs of both subsystems. Three different scenarios were developed, in which distinct recombination rates for genes was used to improve the quality of solutions.

In [25], the researcher model a P-D network over a tactical planning horizon with uncertain demands and capacity. The proposed algorithm incorporates a simulation and an optimisation module; each calculates the total costs of the network for P-D. The problem is mathematically formulated by a MILP, and the fitness function (total cost) is evaluated via the simulation core. Then the solution resulting from the optimisation module is compared with the obtained output from the simulation module recursively. This procedure iterates until there is a set difference between two solutions. This study reports on data obtained from the implementation of the proposed SOA on a SN problem of a reduced scale. Although the simulation and the optimisation modules are both included in the proposed approach, there is no interaction or connection between them. The application of the simulation module is used to produce initial values for the parameters of the mathematical model. Also, the capacity to generalise the model for similar or larger problems was not addressed. Moreover, no evidence was shown around approaching a solution with better quality if different configurations were chosen for the optimisation parameters.

In [15], the authors developed a modified Particle Swarm Optimisation model (MPSO) for a location-allocation Supply Network problem. They formulated a two-echelon Distribution Network (DN) considering multi-product and multi-period inventory, subject to uncertainty of seasonal demands. The determination of the orders quantity and the vendors' location are pursued as the main objectives in this paper. They use Taguchi to tune the parameters of the MPSO. They considered parameter tuning in their model and they performed a sensitivity analysis for similar problems with different granularity levels.

In a similar study, In [26], the researcher developed a PSO algorithm attempting to find the maximum profit for a channel of a two-echelon SN for a single product. Sales quantity and production rate were used as decision variables of their model. Using a combination of GA, PSO, and simulated annealing (SA), they conduct a detailed sensitivity analysis. However, the improvement of the proposed heuristic is computed by using another heuristic. This seems very inefficient.

In [27], the authors proposed a simulation optimisation approach to reduce the number of delayed customer orders while costs are kept under control for an integrated production-distribution supply chain. The hybrid modelling combined linear programming and discrete event simulation. This research is a great potential of using SOA approach; however, no effort was made considering the tunning of the control factors of the GA algorithm.

In [28], the researchers developed an agent-based simulation optimisation model through which an online auction policy within the context of the agricultural supply chain was optimised. Three different scenarios namely, oversupply, balance and insufficient supply with different demand and supply quantities were presented to obtain the optimal lot-size and to determine the optimum online auction policy to control inventory. The investigation towards improving the solution quality derived from the proposed methodology was not provided.

An important observation concerning SOA studies is that, in almost all studies, the tuning of the model's variables (e.g. lead time, production rate, etc.) was only attempted in the optimisation module for small problems. Good examples are included in [20] and [22]. On the other hand, evidence in this regard seems to be missing in some studies [23, 29]. Furthermore, very few ([15, 24]) indicated efforts for tuning the optimisation parameters - selection methodology, mutation, and crossover in GA or swarm's cognitive and social components in PSO. They reported that this had been done by trial and error - a typical approach used in the majority of OR studies [21]. The simulation model is run several times, then the better solution is selected. Due to the complexity of the interaction of parameters of the search algorithm as well as the high computational cost, it is unclear how many iterations would be sufficient for a given size problem. Besides, as the scale of the problem increases, the complexity of interactions increases exponentially. Therefore, the difficulties corresponding to this class of SNP problems will further escalate if a more detailed model is simulated. So, it is necessary to study in more depth the variation of the solution quality.

This paper presents an integrated simulation-optimisation approach to solve a class of CSN problem using GA. The objective is to minimise the total cost while an optimum/near optimum inventory level associated to each product family is obtained. An important feature of the under-investigated problem is that both demand and the inventory capacity are uncertain. The randomness of the uncertain parameters is captured by the simulation model. The optimal quantities are searched by GA. Also, a fine-tunning mechanism for the optimisation algorithm's controllable parameters is applied using Taguchi experimental design and ANOVA to improve the quality of the solution. In Section III, the mathematical model, parameters and notations of the proposed problem are summarised.

3. Mathematical Model

This section presents a mathematical model for a multi-product multi-period consumer supply network. The mathematical model presented here consider a planning period of T (indexed by t), a set of product family P (indexed by i) and a set of retailers R (indexed by j) with the limited budget and inventory restrictions.

The parameters in the model are the following:

D_{ijt}	Demand for product family i by retailer j in period t
D_{ijT}	Demand for product family i by retailer j at the end of period T
I_{0_i}	Initial inventory level for product family i
$O_{\min_{ijt}}$	Minimum quantity of product family i manufactured for retailer j in period t
$O_{\max_{ijt}}$	Maximum quantity of product family i manufactured for retailer j in period t
$V_{\max}$	Maximum capacity of the inventory at DC
V_t	Total capacity of inventory at DC in period t
a_{ijt}	Cost for the ordering of product family i
b_{ijt}	Cost for purchasing one unit of product family i at time t
c_{ijt}	Storage cost for one unit of product family i in period t
d_{ijt}	Handling cost at DC for one unit of product family i in period t
e_{ijt}	Cost for backordering one unit of product family i in period t
f_{ijt}	Cost for transporting one unit of product family i in period t
$\mathcal{A}_T O$	Total cost of ordering at the end of period T
$\mathcal{B}_T I$	Total cost of storage in inventory at the end period T
$\mathcal{C}_T I$	Total cost of handling in inventory at the end of period T
$\mathcal{D}_T D$	Total cost of purchasing at the end of period T
$\mathcal{E}_T O$	Total cost of order shortage at the end period T
$\mathcal{F}_T O$	Total cost of transportation at the end of period T
C_T	The total network costs at the end of period T
σ_1	The backorder intensity rate for product family i at the end of period T
σ_2	The capacity severity rate for product family i at the end of period T

The objective function (1) comprises the minimisation of the total CSN costs, consisting of ordering costs, purchasing costs, transportation costs from manufacturing plants (MP) to retailers (RE), inventory holding and handling costs at the distribution centre (DC), and backordering costs subject to a set of constraints present in (2-4). Constraint (1) represents the quantity of order of a product family i in a period t bounded by the upper and the lower limits. Note, the maximum quantity of an order for product family i from retailer j cannot exceed maximum n folds of the maximum quantity of the demand for the entire planning period T. Constrain (2) is the capacity of the inventory denoted by V_T. The order quantity is a positive integer that is normalised between 0 and 1 by (4) denoted by $\acute{O}$. Table 1 and Table 2 shows a numerical representation of $O_{ijt}, \acute{O_{ijt}}$ and D_{ijt} for $i = 3, j = 5$ and $t = 2$.

$$min \sum_{t=1}^{T}\sum_{j=1}^{R}\sum_{i=1}^{P} C_{ijt}(O_{ijt}, I_{ijt}) \tag{1}$$

$$C_T(O_{ijt}, I_{ijt}) = \mathcal{A}_T(O_{ijt}) + \mathcal{B}_T(I_{ijt}) + \mathcal{C}_T(I_{ijt}) + \mathcal{D}_T(D_{ijt}) + \mathcal{E}_T(I_{ijt}) + \mathcal{F}_T(O_{ijt}) \quad ; \forall i,j \geq 0$$

$$\begin{cases} \mathcal{A}_T(O_{ijt}) = a_{ijt}.O_{ijt} \\ \mathcal{B}_T(I_{ijt}) = b_{ijt}.I_{ijt} \\ \mathcal{C}_T(I_{ijt}) = c_{ijt}.I_{ijt} \end{cases} \quad \begin{cases} \mathcal{D}_T(D_{ijt}) = d_{ijt}.D_{ijt} \\ \mathcal{E}_T(I_{ijt}) = e_{ijt}.I_{ijt} \\ \mathcal{F}_T(O_{ijt}) = f_{ijt}.O_{ijt} \end{cases}$$

$$C_T(O_{ijt}, I_{ijt}) = minimise \sum_{t=1}^{T}\sum_{j=1}^{R}\sum_{i=1}^{P} a_{ijt}.O_{ijt} + b_{ijt}.I_{ijt} + c_{ijt}.I_{ijt} + d_{ijt}.D_{ijt} + e_{ijt}.I_{ijt} + f_{ijt}.O_{ijt}$$

subject to:

$$O_{min} \leq O_{ijt} \leq O_{max} \tag{2}$$

$$O_{min}, O_{max} = \begin{bmatrix} 0 & n * \max(D_{ijT}) \end{bmatrix} \quad ; \quad n > 1 \tag{3}$$
$$V_{max} \leq V_T$$

$$O_{ijt} = \min\left(\left\lceil Omin + (Omax - Omin + 1) \times \acute{O} \right\rceil, Omax\right) \tag{4}$$
$$0 \leq \acute{O} \leq 1$$

Table 1. Numerical representation of $\acute{O_{ijt}}, O_{ijt}$

$\acute{O}_{ijt}$	$\acute{O}_{11t}$	$\acute{O}_{12t}$	$\acute{O}_{13t}$	...	$\acute{O}_{1jt}$
	$\acute{O}_{21t}$	$\acute{O}_{22t}$	$\acute{O}_{23t}$	...	$\acute{O}_{2jt}$
	⋮	⋮	⋮	⋮	⋮
	$\acute{O}_{i1t}$	$\acute{O}_{i2t}$	$\acute{O}_{i3t}$	...	$\acute{O}_{ijt}$
$\acute{O}_{ij1}$	0.771	0.134	0.681	0.414	0.820
	0.699	0.568	0.332	0.247	0.962
	0.697	0.425	0.106	0.929	0.581
$\acute{O}_{ij2}$	0.338	0.040	0.182	0.887	0.991
	0.670	0.306	0.771	0.135	0.092
	0.017	0.394	0.973	0.116	0.447
O_{ijt}	O_{11t}	O_{12t}	O_{13t}	...	O_{1jt}
	O_{21t}	O_{22t}	O_{23t}	...	O_{2jt}
	⋮	⋮	⋮	⋮	⋮
	O_{i1t}	O_{i2t}	O_{i3t}	...	O_{ijt}
O_{ij1}	161	240	36	111	57
	176	172	52	145	231
	103	96	97	58	56
O_{ij2}	259	248	53	237	288
	212	158	87	124	68
	110	99	196	164	132

Table 2. numerical representation of D_{ijt}

D_{ijt}	D_{11t}	D_{12t}	D_{13t}	...	D_{1jt}
	D_{21t}	D_{22t}	D_{23t}	...	D_{2jt}
	⋮	⋮	⋮	⋮	⋮
	D_{i1t}	D_{i2t}	D_{i3t}	...	D_{ijt}
D_{ij1}	259	248	53	237	288
	212	158	87	124	68
	110	99	196	164	132
D_{ij2}	11	26	17	35	38
	72	93	80	42	6
	69	61	87	39	85

Note: D_{11t} presents the quantity of product family 1 to be manufactured for consumer 1 in time interval $t = 1$ is 259 unit.

The I_{ijt} and O_{ijt} are related to the decisions regarding the inventory level and the quantity of orders that are calculated by (5). O_{ijt} is the main decision variable, since I_{ijt} is obtained recursively from O_{ijt}. The demand quantity, D_{ijt}, is unknown but bounded. It can be expressed by probabilistic distribution functions such as normal or uniform distribution functions. In this model, a uniform distribution is used to model D_{ijt} using (6), where D_{min}, D_{max} are the lower and upper bounds, respectively.

Also, each product family has a set volume (v_i) so the total volume of the order i.e. the total volume occupied by the inventory, V_{max} , is calculated by (7)

$$O_{ijt} = I_{ijt} - I_{ijt-1} + D_{ijt} \tag{5}$$

$$D_{ijt} \sim U\,(D_{min}\,, D_{max}) \tag{6}$$

$$V_{max} = \sum_{i=1}^{P}\sum_{j=1}^{R}\sum_{t=1}^{H} v_i \times I_{ijt} \qquad v_i \sim U\,(0,1) \tag{7}$$

If a solution breaks any constraint (c_i) it is infeasible and therefore the associated evaluation should be penalised in proportion to how violently they break the constraints. In this problem $\propto_1$ and $\propto_2$ are defined and assigned to the fitness function via (8). The problem size and substantially the changes in the planning period result in changes of $\propto_1, \propto_2$.

$$C_T(O_{ijt}, I_{ijt}) = \left\{\left(\sum_{t=1}^{T}\sum_{j}^{R}\sum_{i=1}^{P} \mathcal{A}_T.O_{ijt} + \mathcal{B}_T.I_{ijt} + \mathcal{C}_T.I_{ijt} + \mathfrak{D}_T.D_{ijt} + \mathcal{E}_T.I_{ijt} + \mathcal{F}_T.O_{ijt}\right) + \alpha\,\sigma_1\right\} \times (1 + \beta\,\sigma_2) \tag{8}$$

$$\sigma_1 = \frac{\sum_t^T \sum_j^R \sum_i^P I_{ijt}}{TRP}$$

$$I_{ijk} < 0$$

$$\sigma_2 = \frac{\left(V_{max} - \frac{\sum_t^T V_t}{V_{max}}\right)}{T}$$

Also, the average backlogged orders, and the average volume occupied by the inventory are denoted by σ_1 and σ_2, respectively. In associate with the planning policy in-use, the values of $\sigma_1\,, \sigma_2$ may vary. For example, if the customer satisfaction rate is %100, which means shortages are not allowed and $\sigma_1 = 0$. Conversely, if a company unable to deliver their promises on time then σ_1 can be set according to the safety stock level. Note, in both cases, the inventory capacity cannot be exceeded, thus $\sigma_2 = 0$. So, a solution candidate is regarded feasible if both conditions are satisfied.

4. Granularity

In systems engineering literature, granularity translates into the level of detail one can decide to consider in a model or decision-making process where the same functionality is expressed with different 'sized' designs [30]. In SN, the size of the problem determines the granularity level of the problem which has a significant influence on the computation time and the algorithm's efficiency. Measures such as the number of product families, the number of facilities, planning periods, etc. are some important factors which affect the granularity level [31]. In this study, in order to verify the robustness of the proposed methodology, three case studies with different granularity levels are considered for the design of experiments represented by a tree structure with two levels L_1 and L_2 (Figure 2). The leaves at L_1 denoted by $[P_S, P_M, P_L]$, correspond to an individual scenario with a distinct problem size, known as *Small*, *Medium* and *Large-scale* problems. L_1 is developed based on the problem size categories proposed by Mousavi, Bahreininejad, Musa and Yusof [15], shown in Table 3. The roots at L_2 are the number of experiments considered for each category. This is determined according to the number of parameters and the levels of variation of a specific parameter which will be developed using Taguchi method (see Section 6).

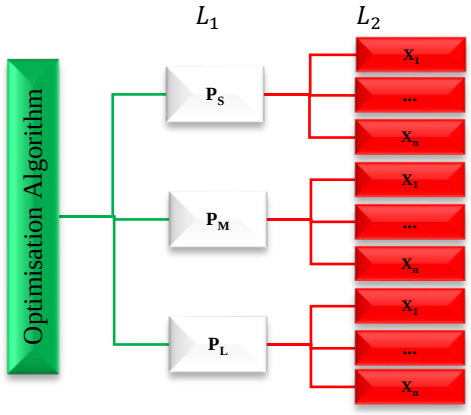


Figure 2 Hierarchical structure proposed for implementation phase

Table 3. Sizes of the proposed instances [15]

Problem Size	Product Family (P)	Manufacturing Plants (MP)	Retailer (RE)	Periods (T)
Small	[1-5]	[1-5]	[5-10]	[1-3]
Medium	[6-10]	[1-10]	[11-20]	[1-5]
Large	[11-15]	[11-15]	[20-30]	[6-10]

Note: a problem with $P = 7, MP = 6, RE = 11$ and $T = 2$ is counted as a Medium-scale problem.

5. Solution Approach

To solve the MPMPCSN problem discussed in this paper, GA optimisation method is used. GA are based on principles of natural selection and genetics to evolve better solutions through multiple consecutive generations. *Selection*, *Crossover* and *Mutation* are implementations in GA of similar phenomena occurring in the natural world. [23]. Based on the quality of solutions, they have a probability to be selected and evolve in new generations and converge towards optimality. Finally, the solutions are tested against termination criteria (evolving procedure). A good search space and genetic operators must maintain an equilibrium between exploration and exploitation and this is key in reaching optimality [32-34]

5.1. Generation and Initialisation

The first step in implementing the GA is to generate a random population of solutions (chromosomes). Chromosomes are resizable according to problem's attributes and vary based on the problem type, level of complexity, number and type of variables, granularity, etc. Each chromosome consists of several atomic structures - *genes* representing the characteristics of the solution (e.g. number of suppliers, position of manufacturing plants, types of products considered, etc.) [35]. Real coding has been used for this type of problem (Figure 3).

1	$\acute{O}_{ijt}$	$\acute{O}_{11t}$	$\acute{O}_{12t}$	$\acute{O}_{13t}$	...	$\acute{O}_{1jt}$	TC_{1t}
2		$\acute{O}_{21t}$	$\acute{O}_{22t}$	$\acute{O}_{23t}$	...	$\acute{O}_{2jt}$	TC_{2t}
⋮		⋮	⋮	⋮	⋮	⋮	⋮
n_{Pop}		$\acute{O}_{i1t}$	$\acute{O}_{i2t}$	$\acute{O}_{i3t}$	...	$\acute{O}_{ijt}$	$TC_{n_{Pop}}$

Figure 3 Chromosome representation

The performance of the GA is affected by two opposing factors; population size and computation time. The larger the population size; the longer takes the computation time. The population size should be large enough to incorporate sufficient variation in one generation from which the children in the next

generations are produced. GA is designed to evolve over a number of generations. Hence, having a large population has a serious impact on the computation time. A carefully selected population size that offers sufficient variety but does permit a fast-enough evolution is needed.

5.2. Genetic Operators: Selection, Crossover, and Mutation

Genetic operators may affect the optimal fitness value for the designed algorithm. The GA operators presented in this paper are selection, crossover and mutation. Roulette Wheel, Tournament and Ranked are the most popular selection mechanisms that are used in this study [33, 36].

In the following step, the offspring population is created by applying single point crossover and mutation. So, new offsprings are produced by combining the characteristics of two parents that can be better than both parents if they take the best characteristics from each of the parents. This mechanism should be performed with a certain probability. Throughout this study, P_c and P_m are refferring to crossover and mutation probabilities respectively. Two individuals are produced per randomly selected parents followed by mutating gens of offspring population with specified probability. The mutation is implemented to preserve the variety of the solution pool and prevent GA getting stuck in local optima by exploring the entire search space and maintaining diversity in the population [37]. It is likely that some randomly lost genetic information recovered through mutation. P_m should be set carefully too as such the diversity in the population is preserved but does not negatively affect the overall, fitness of the current population by removing good solutions. Mutation can finely tune the balance between exploration and exploitation. Typically, the mutation rate is small (<2-5%).

5.3. Simulation

After initialising the first population, each chromosome is evaluated for fitness. Fitness function is a metric used to measure the quality of the represented solution. The fitness value of a chromosome is the most important factor in GA evaluation that is always problem dependent [38]. The fitness function defined for MPMPCSN is the minimum cost of running the network. So the lower the fitness value, the higher is the survival chance of a chromosome.

5.4. Stopping Conditions

The optimal/near optimal solution is achieved through an iterative procedure until the stopping condition is satisfied. Choosing the termination criteria depends on the complexity of the problem structure as well as the size of the solution pool [39]. Often, the maximum number of generation is adopted which is the case in this study.

The traditional GA has several shortcomings. As a result of premature convergence, the search parameters (selection, crossover, mutation) may not be very useful towards the end of a search procedure [40]. Also, obtaining an absolute global optimum is not guaranteed, however providing good solutions within a reasonable time is generally expected [41, 42]. Also, GA may not be effective if the starting point in search space was at a great distance from optimal solutions [43]. This deficiency limits the use of GA in real-time applications. However, it can be overcome if GA is hybridised with other local search methods where a closed-form expression of the objective function can be appropriately performed [42]. Simulation tools are unique methods that are tightly integrated with mathematical and algorithmic based models. Overall, to improve GA performance and obtain accurate solutions, the population size, selection mechanism, crossover and mutation rates and the computational time are required to be turned. Further validation and evaluation of the proposed model and the solution approach is discussed in the following section.

6. Computational Experiments

This section provides experimental results obtained from applying the proposed SOA methodology on practical tests associated to MPMPCSN problems with different granularity levels.

A manufacturing CSN with a central distribution centre is considered in which orders received from consumers are being processed. The demand quantities for P_S , P_M and P_L were randomly generated first and remained unchanged throughout the rest of the optimisation algorithm (see. Appendix A, Table 22-Table 24), because the variation of D_{ijt} causes changes of other parameters. Also, associated purchase cost per unit of product family P_i and the corresponding volume v_i for P_S, P_M and P_L are given in Appendix A (Table 21). All other related costs of running the network consist of ordering cost, backordering cost, holding cost, handling cost and transportation cost are computed via (9)-(14). In addition, the fixed parameters of the model are presented in Table 4.

$$a_{ijt} = 0.1 \times d_{ijt} \quad (9)$$

$$b_{ijt} = 0.05 \times d_{ijt} \quad (10)$$

$$c_{ijt} = 0.05 \times d_{ijt} \quad (11)$$

$$1 \leq d_{ijt} \leq 100 \quad (12)$$

$$e_{ijt} = 0.05 \times d_{ijt} \quad (13)$$

$$f_{ijt} = 0.05 \times d_{ijt} \quad (14)$$

Table 4. fixed parameters of the model

Parameters	P	R	T	V_T
Small-Scale	5	2	2	1000
Medium-Scale	6	11	5	10000
Large-Scale	10	25	8	100000

Note: P, R, and T are referred to the Product family, Retailer and Planning period respectively.

7. Results and Discussion

As discussed above, the performance of the GA optimisation algorithm is mostly influenced by its controllable parameters. These parameters are selection method (P_s), crossover and mutation rate (P_c, P_m), population size (n_{Pop}) and the maximum number of iteration ($MaxIt$). Thus, though utilising Taguchi Orthogonal Array Design along with Regression Analysis and Optimisation Solver the optimal parameter set was determined. More details are given in the following sections.

7.1. Process of Experiment Design

The main two components of the Taguchi method are the number of parameters and their variation levels. In order to analyse the results obtained from ANOVA (analysis of variance) and S/N ratio (signal to noise), it is required to create a set of tables of

numbers known as *orthogonal arrays.* These tables are then used first to reduce the number of experiments, next to determine the most critical parameters with high impact on the outcomes. In this study, we consider the GA controllable parameters as significant factors in 3 levels (Table 7). The Taguchi Orthogonal Array Design ($L27 - 3^5$) shown in Table 6 is proposed and created by Minitab.

Table 5. The *GA* parameters' level

Granularity Level	Small-scale	Medium-Scale	Large-Scale
Parameters	Level 1	Level 2	Level 3
P_s	*RW*	*T*	*R*
P_c	0.9	0.85	0.8
P_m	0.1	0.05	0.025
n_{Pop}	[30 60 120]	[100 150 200]	[100 200 300]
$MaxIt$	[200 100 50]	[500 400 300]	[3500 3000 2000]

RW, *T* and *R* referred to Roulette Wheel, Tournament and Ranked Selection method respectively

Table 6. The layout of the orthogonal array for 5 factors in 3 levels

No.	P_s	P_c	P_m	n_{Pop}	$MaxIt$
S1	1	1	1	1	1
S2	1	1	1	1	2
S3	1	1	1	1	3
S4	1	2	2	2	1
S5	1	2	2	2	2
S6	1	2	2	2	3
S7	1	3	3	3	1
S8	1	3	3	3	2
S9	1	3	3	3	3
S10	2	1	2	3	1
S11	2	1	2	3	2
S12	2	1	2	3	3
S13	2	2	3	1	1
S14	2	2	3	1	2
S15	2	2	3	1	3
S16	2	3	1	2	1
S17	2	3	1	2	2
S18	2	3	1	2	3
S19	3	1	3	2	1
S20	3	1	3	2	2
S21	3	1	3	2	3
S22	3	2	1	3	1
S23	3	2	1	3	2
S24	3	2	1	3	3
S25	3	3	2	1	1
S26	3	3	2	1	2
S27	3	3	2	1	3

7.2. Signal-to-Noise (S/N) Ratio Method

S/N ratios evaluate the size of the apparent effect (signal) against the size of random fluctuations (noise) witnessed in the data. The higher this indicator, the better the compromise is which can be calculated in different ways according to the optimisation problem (minimisation/maximisation) [44]. In this study, S/N ratio values are calculated to determine the best combination of GA control factors. The proposed optimisation algorithm was run four times for each parameter set to obtain more refined solutions. The numerical results for the Small, Medium and Large-scale problem are reported in Table 7, Table 8 and Table 9, respectively.

This problem is aimed to minimise the response value (y). Therefore, to minimise the mean-square deviation (MSD) from the target value 0 and maximise the S/N ratio, MSD has to be calculated using (15). The signal to noise (S/N) ratio, in this case, is defined by (16), where n is the sample size.

Table 7. Taguchi experimental design and design data of *GA* for small-scale problem

Trial	Function Evaluation (TC)				μ	σ	P_s	P_c	P_m	n_{Pop}	$MaxIt$
	Run 1	Run 2	Run 3	Run 4							
1	52545.31	52838.97	52824.62	52798.99	52751.97	138.76	RW	0.9	0.1	30	200
2	57736.13	57984.64	54800.67	56440.22	56740.41	1459.71	RW	0.9	0.1	30	100
3	55082.79	54767.13	55334.41	55983.30	55291.91	516.06	RW	0.9	0.1	30	50
4	57348.28	56895.83	58086.99	58118.95	57612.51	595.84	RW	0.85	0.05	60	200
5	59594.91	58612.27	61314.77	61253.54	60193.87	1321.56	RW	0.85	0.05	60	100
6	60380.16	62646.26	60710.87	59366.54	60775.96	1371.79	RW	0.85	0.05	60	50
7	55536.78	54608.65	55060.74	54506.04	54928.05	471.97	RW	0.8	0.025	120	200
8	55135.85	54540.19	54946.94	56517.07	55285.01	858.15	RW	0.8	0.025	120	100
9	57518.01	59179.99	56537.55	57925.29	57790.21	1094.37	RW	0.8	0.025	120	50
10	52410.97	52718.79	52428.90	52416.20	52493.72	150.24	T	0.9	0.05	120	200
11	53368.53	52881.84	53767.00	52857.57	53218.73	434.73	T	0.9	0.05	120	100
12	58698.26	55432.41	56344.90	57940.46	57104.01	1484.56	T	0.9	0.05	120	50
13	54263.36	56283.66	55064.54	55837.51	55362.27	889.02	T	0.85	0.025	30	200
14	56139.17	56388.68	57656.13	56204.80	56597.20	713.81	T	0.85	0.025	30	100
15	62448.49	94741.69	60631.15	98432.34	79063.42	20304.09	T	0.85	0.025	30	50
16	52413.82	52417.87	52439.46	52418.27	52422.36	11.58	T	0.8	0.1	60	200
17	53546.80	54432.52	53665.56	52804.39	53612.32	666.48	T	0.8	0.1	60	100
18	62686.18	56408.68	56602.45	56552.31	58062.41	3083.61	T	0.8	0.1	60	50
19	54034.56	53650.51	53214.02	53760.76	53664.96	341.24	R	0.9	0.025	60	200
20	56947.05	58519.69	57332.37	56946.45	57436.39	744.73	R	0.9	0.025	60	100
21	62368.65	58889.81	64213.45	64114.90	62396.70	2486.75	R	0.9	0.025	60	50
22	52472.93	52454.69	52466.57	52462.89	52464.27	7.61	R	0.85	0.1	120	200
23	54151.02	54381.73	54913.21	54443.82	54472.45	319.71	R	0.85	0.1	120	100
24	59054.53	58677.67	59390.45	59848.09	59242.69	497.66	R	0.85	0.1	120	50
25	54123.74	53139.69	53600.31	53588.71	53613.11	402.34	R	0.8	0.05	30	200
26	62582.39	57133.15	57636.73	58226.32	58894.65	2498.75	R	0.8	0.05	30	100
27	76782.74	63219.40	67855.77	65419.86	68319.44	5951.48	R	0.8	0.05	30	50

Note: (μ: m ean , σ: St andard deviation)

Table 8. Taguchi experimental design and design data of *GA* for medium-scale problem

Trial	Function Evaluation (TC)				μ	σ	P_s	P_c	P_m	n_{Pop}	$MaxIt$
	Run 1	Run 2	Run 3	Run 4							
1	3055303	3053526	3046047	3050184	3051265.00	4074.87	RW	0.9	0.1	200	500
2	3149794	3154852	3180213	3164676	3162383.75	13396.09	RW	0.9	0.1	200	400
3	3372901	3350114	3335613	3323874	3345625.50	21114.59	RW	0.9	0.1	200	300
4	3200575	3185842	3197118	3191536	3193767.75	6464.29	RW	0.85	0.05	150	500
5	3355893	3308538	3369709	3382514	3354163.50	32301.14	RW	0.85	0.05	150	400
6	3499418	3511169	3529597	3529401	3517396.25	14775.75	RW	0.85	0.05	150	300
7	3432256	3440475	3410509	3433997	3429309.25	13022.82	RW	0.8	0.025	100	500
8	3575145	3520148	3586398	3537586	3554819.25	31141.79	RW	0.8	0.025	100	400
9	4555883	4146796	3846552	4203898	4188282.25	290903.67	RW	0.8	0.025	100	300
10	3051447	3066724	3034552	3045986	3049677.25	13368.15	T	0.9	0.05	100	500
11	3156857	3217344	3129544	3179152	3170724.25	37114.87	T	0.9	0.05	100	400
12	3281164	3310920	3406627	3340245	3334739.00	53652.67	T	0.9	0.05	100	300
13	3077422	3072374	3047223	3078703	3068930.50	14727.31	T	0.85	0.025	200	500
14	3182456	3188477	3166677	3221685	3189823.75	23144.54	T	0.85	0.025	200	400
15	3436084	3441521	3417875	3435688	3432792.00	10294.60	T	0.85	0.025	200	300
16	2991777	2972519	2986549	2982617	2983365.50	8146.47	T	0.8	0.1	150	500
17	3057430	3030744	3064818	3033992	3046746.00	16926.05	T	0.8	0.1	150	400
18	3172227	3184862	3188181	3173263	3179633.25	8079.53	T	0.8	0.1	150	300
19	3360788	3373308	3403272	3440016	3394346.00	35280.59	R	0.9	0.025	150	500
20	3503662	3492818	3501245	3457735	3488865.00	21267.49	R	0.9	0.025	150	400
21	6083231	6707308	6357912	5970323	6279693.50	328268.14	R	0.9	0.025	150	300
22	3099402	3117656	3130297	3111689	3114761.00	12846.33	R	0.85	0.1	100	500
23	3243754	3249067	3272814	3255208	3255210.75	12634.31	R	0.85	0.1	100	400
24	3410574	3462829	3421737	3409948	3426272.00	24965.85	R	0.85	0.1	100	300
25	3232477	3281042	3288839	3245727	3262021.25	27198.93	R	0.8	0.05	200	500
26	3372780	3354390	3375793	3360978	3365985.25	10031.33	R	0.8	0.05	200	400
27	3542951	3526380	3567064	3540808	3544300.75	16865.55	R	0.8	0.05	200	300

Note: (μ: m ean , σ: St andard deviation)

Table 9. Taguchi experimental design and design data of GA for large-scale problem

Trial	Function Evaluation (TC)				μ	σ	P_s	P_c	P_m	n_{Pop}	$MaxIt$
	Run 1	Run 2	Run 3	Run 4							
1	6197853	6182641	6205040	6171968	6189375	14895.51	RW	0.9	0.1	100	3500
2	6171968	6197853	6182641	6205040	6189375	14895.51	RW	0.9	0.1	100	3000
3	6026883	6036349	6064389	6092117	6054934	29462.9	RW	0.9	0.1	100	2000
4	6171329	6156044	6162801	6148266	6159610	9813.874	RW	0.85	0.05	200	3500
5	6189910	6160183	6168566	6183770	6175607	13646.61	RW	0.85	0.05	200	3000
6	6276588	6197853	6256788	6232034	6240816	33949.63	RW	0.85	0.05	200	2000
7	5609430	5583614	5604952	5587145	5596285	12806.33	RW	0.8	0.025	300	3500
8	6219773	6220941	6220798	6296291	6239450	37896.94	RW	0.8	0.025	300	3000
9	6393839	6421235	6397965	6500502	6428385	49567.4	RW	0.8	0.025	300	2000
10	5765313	5783485	5797242	5786545	5783146	13270.95	T	0.9	0.05	300	3500
11	6145198	6115210	6141100	6146131	6136910	14630.8	T	0.9	0.05	300	3000
12	6181667	6166755	6174604	6150186	6168303	13526.68	T	0.9	0.05	300	2000
13	5766122	5797580	5768570	5819409	5787920	25393.04	T	0.85	0.025	100	3500
14	6330538	6295429	6350012	6405480	6345365	46003.23	T	0.85	0.025	100	3000
15	6421425	6446814	6429805	6425072	6430779	11227.04	T	0.85	0.025	100	2000
16	6124129	6234488	6150018	6149727	6164591	48152.91	T	0.8	0.1	200	3500
17	6132648	6141044	6166400	6151393	6147871	14537.94	T	0.8	0.1	200	3000
18	5803931	5783648	5803967	5805327	5799218	10400.52	T	0.8	0.1	200	2000
19	5930494	5953702	5898563	5878441	5915300	33387.83	R	0.9	0.025	200	3500
20	6231820	6227294	6232032	6276543	6241922	23183.89	R	0.9	0.025	200	3000
21	6401559	6416416	6425043	6414352	6414342	9699.475	R	0.9	0.025	200	2000
22	6103190	6123392	6079358	6102566	6102126	17999.54	R	0.85	0.1	300	3500
23	5729010	5873701	5867909	5715137	5796439	86089.04	R	0.85	0.1	300	3000
24	6103190	6123392	6079358	6122019	6106989	20598.34	R	0.85	0.1	300	2000
25	6271088	6235294	6240251	6212358	6239748	24169.66	R	0.8	0.05	100	3500
26	6219361	6297088	6249893	6228781	6248781	34642.66	R	0.8	0.05	100	3000
27	6402903	6433575	6407674	6422437	6416647	14017.7	R	0.8	0.05	100	2000

Note: ($\boldsymbol{\mu}$: m ean , $\boldsymbol{\sigma}$: St andard deviation)

$$MSD = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \tag{15}$$

$$\frac{S}{N} = -10\log(MSD) \tag{16}$$

The example of the calculation of S/N ratio for the control parameter P_s is shown below (column 1 of Table 10) and the results correspond to each case study are summarised in Table 10, Table 11 and Table 12. The difference between the levels of factors in the Table 10- Table 12 determines which parameter has more effect on the quality characteristics (the total cost of the network).

$$Level\ 1 = \frac{(-94.44 - 95.08 - 94.85 - 95.21 - 95.59 - 95.68 - 94.80 - 94.85 - 94.24)}{9} = -95.08$$

$$Level\ 2 = \frac{(-94.40 - 94.52 - 95.13 - 94.86 - 95.05 - 98.16 - 94.39 - 94.58 - 95.28)}{9} = -95.16$$

$$Level\ 3 = \frac{(-94.60 - 95.18 - 95.91 - 94.40 - 94.the\ 72 - 95.45 - 94.59 - 95.41 - 96.72)}{9} = -95.22$$

$$Difference = |highest\ value| - |lowest\ value| = |-95.22| - |-95.08| = 0.14$$

As it can be seen from Table 10, the control factor $MaxIt$, by far is the most important factor that impacts on S/N ratio (1.19), n_{Pop} , P_m , P_c and P_s are also significant factors. Table 11 shows $MaxIt$, P_s and P_m are approximately double of P_c and n_{Pop}. Also, in Table 12 while control factor P_c has a negligible effect in influencing the S/N ratio in P_L problem, the contribution of all other four parameters (P_s, P_m, n_{Pop} and $MaxIt$) to the S/N is more than 10%.

The S/N ratios computed for the data set P_s, P_c, P_m, n_{Pop} and $MaxIt$ (Table 10-Table 12) are essential for sketching the S/N ratio response diagrams for P_S, P_M and P_L problems (0). So, a higher S/N ratio is related to a data set with the minimum variation which is considered as the best data set.

Table 10. The response table of S/N ratio of P_S Problem

	Selection (P_s)	Crossover Rate (P_c)	Mutation Rate (P_m)	Population Size (n_{Pop})	Generation ($MaxIt$)
Level 1	-95.08	-94.90	-94.80	-95.46	-94.63
Level 2	-95.16	-95.46	-95.25	-95.16	-95.00
Level 3	-95.22	-95.10	-95.41	-94.84	-95.83
Difference	0.14	0.56	0.61	0.63	1.19

Table 11. The response table of S/N ratio P_M Problem

	Selection (P_s)	Crossover Rate (P_c)	Mutation Rate (P_m)	Population Size (n_{Pop})	Generation ($MaxIt$)
Level 1	-130.7	-130.9	-130	-130.3	-130
Level 2	-130	-130.3	-130.4	-130.9	-130.3
Level 3	-131.1	-130.6	-131.3	-130.6	-131.4
Difference	1.1	0.5	1.3	0.6	1.4

Table 12. The response table of S/N ratio P_L Problem

	Selection (P_s)	Crossover Rate (P_c)	Mutation Rate (P_m)	Population Size (n_{Pop})	Generation ($MaxIt$)
Level 1	-135.8	-135.6	-135.6	-135.9	-135.5
Level 2	-135.7	-135.7	-135.8	-135.8	-135.8
Level 3	-135.8	-135.8	-135.7	-135.6	-135.9
Difference	0.1	0.2	0.2	0.3	0.4

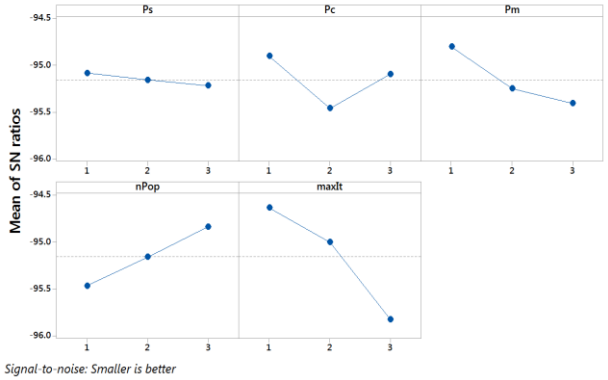
Therefore, the best values associated with P_s, P_c, P_m, n_{Pop} and $MaxIt$ corresponding to P_S, P_M and P_L problems are as follows: for $\boldsymbol{P_S}$, level 1(Roulette Wheel selection), level 1 (90% crossover), level 1 (10% mutation), level 2 (120 chromosomes) and level 1 (200 iterations), respectively; for $\boldsymbol{P_M}$ level 2 (Tournament selection), level 2 (85% crossover), level 1 (10% mutation), level 1 (200 chromosomes) and level 1 (500 iterations), respectively; For $\boldsymbol{P_L}$ level 2 (Tournament selection), level 1 (90% crossover), level 1 (10% mutation), level 3 (300 chromosomes) and level 1 (3500 iterations), respectively. This can be observed from S/N ratio response diagrams too (Figure 4). The rows show difference values in Table 10-Table 12 determine the contribution level of each parameter in obtaining lower cost. So, the total cost of running the network, for example for P_M problem, is mostly affected by the number of generation, mutation rate, the selection method, population size and crossover rates of the GA algorithm. To determine the significant level of these parameters, ANOVA method is utilised for which the data given in Table 7- Table 9 are going to be used again. Results obtained from ANOVA are summarised in Table 13-Table 15.

7.3. ANOVA Method

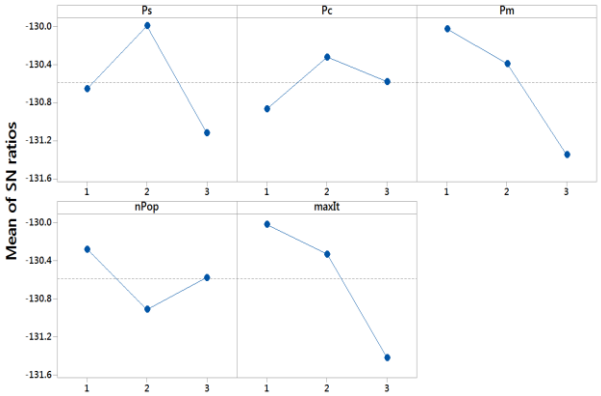
From ANOVA, the percentage contribution ratio (PCR) of each parameter can be calculated. PCR indicates the significance of all main factors and their interactions on the output. The calculation is performed by comparing the mean square (MS) against an estimate of the experimental errors at specific confidence levels. The total sum of squared deviations (SS_T) from the total mean S/N ratio is calculated via (17).

$$SS_T = \sum_{i=1}^{n} (\eta_i - n_m)^2 \tag{17}$$

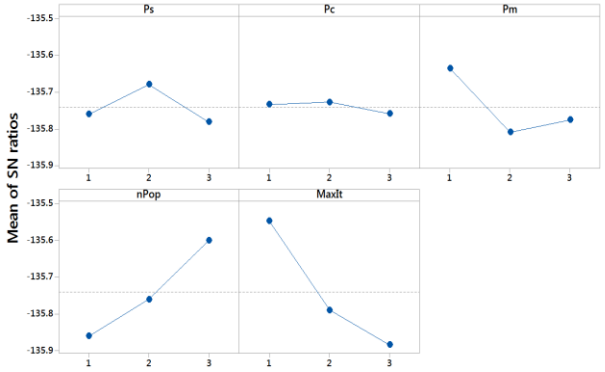
where n is the number of experiments in the orthogonal array and η_i is the mean S/N ratio for the ith experiment.



Small-scale problem



Medium-scale problem



Large-scale problem

Figure 4 The main effect diagram for S/N Ratio response diagram for GA parameters ($P_s, P_c, P_m, n_{Pop}, MaxIt$)

The ANOVA tables for S/N ratios corresponding to the data in Table 10-Table 12 are summarised in Table 13- Table 15. The terms SS_T and SM_T are corresponding to the total sum of squared and the total mean square, respectively. Also, the F-ratios and P-values provided in "F" and "P" columns are calculated via (18) and (19), respectively. F-ratio indicates which parameter ($P_s, P_c, P_m, maxIt$) have a significant effect on the quality characteristic (TC) and P-value determines the significant percentage of the parameters on the quality characteristic (TC).

$$F = \frac{SS_T}{MS_T} \tag{18}$$

$$P = \frac{SS_T}{SS_T} \tag{19}$$

Table 13. Results obtained from ANOVA for Small-scale problem

Source	DF	SS_T	MS_T	F	P
P_s	2	19727169	9863585	0.38	0.685
P_c	2	2.76E+08	1.38E+08	5.32	0.006
P_m	2	3.33E+08	1.66E+08	6.41	0.002
n_{Pop}	2	3.49E+08	1.75E+08	6.72	0.002
$MaxIt$	2	1.24E+09	6.22E+08	23.96	0
Error	97	2.52E+09	25971697		

Table 14. Results obtained from ANOVA for medium-scale problem

Source	DF	SS_T	MS_T	F	P
P_s	2	4.86E+12	2.43E+12	14.27	0
P_c	2	1.69E+12	8.44E+11	4.96	0.009
P_m	2	7.30E+12	3.65E+12	21.45	0
n_{Pop}	2	2.07E+12	1.03E+12	6.08	0.003
$MaxIt$	2	8.19E+12	4.10E+12	24.08	0
Error	97	1.65E+13	1.70E+11		

Table 15. Results obtained from ANOVA for large-scale problem

Source	DF	SS_T	MS_T	F	P
P_s	2	1.02E+11	5.1E+10	1.52	0.223
P_c	2	1.2E+10	5.98E+09	0.18	0.837
P_m	2	3.09E+11	1.55E+11	4.61	0.012
n_{Pop}	2	5.93E+11	2.96E+11	8.85	0
$MaxIt$	2	1.06E+12	5.30E+11	15.82	0
Error	97	3.25E+12	3.35E+10		

Note: SS and V stand for the sum of squared and the variance respectively.

It can be observed from Table 13 that the difference between the mean values of the level of the control factor P_s (selection method) is insignificant ($0.68 > \alpha = 0.05$). Therefore, any selection strategy can be chosen for implementation of the proposed SOA for small-scale problem. However, the difference between the mean values of crossover rates (P_c), mutation rate (P_m) and the number of iteration ($MaxIt$) is significant (0.006, 0.002 and $0.002 < \alpha = 0.05$). Thus, the best control factor setting for maximising the S/N ratio is P_c at level 1, P_m at level 1, n_{Pop} at level 2 and $MaxIt$ at level 1. In the Medium-scale problem, all of the control factors are highly contributing to the performance of the SOA (Table 14). According to Table 15, only P_m, n_{Pop} and $MaxIt$ are significantly influenced on the performance of the SOA in Large-scale problem, while there is no restriction in choosing the selection strategy and the crossover rate.

7.4. Confirmation test

The final step of the verification phase is to perform the confirmation test with the optimal level of the GA parameters drawn based on the Taguchi's design approach for each case study (Table 16).

Table 16. The best combination of the GA parameters

	P_S	P_c	P_m	$nPop$	$MaxIt$
Small-Scale	R	0.9	0.1	120	200
Medium-Scale	T	0.85	0.1	200	500
Large-Scale	R	0.9	0.05	100	3500

The results obtained from the proposed methodology and GA solver associated with P_S, P_M and P_L problems along with the average of the best and the worst results are summarised in Table 17. The quality measurement of the solution is determined according to the value of standard deviation (σ). Therefore, the solution candidate with the maximum σ is considered as the worst solution and the one with the minimum value is regarded as the

best solution. Hence, the experiments No. 15 and No. 22 are the worst and the best scenario for the Small-scale problem, respectively.

Table 17. The total optimised cost

Problem Size	Small-scale	Medium-scale	Large-scale
Optimal Scenario	49966.28($)	2921429.2($)	5971604 ($)
Best Scenario	52464.27($)	3051265($)	6102126 ($)
Worst Scenario	79063.41($)	6279694($)	6239450 ($)

As can be seen from Figure 5, the proposed algorithm shows better performance compared with the best and the worst solutions acquired from GA solver (5% $\cong$ \$ 2498). A similar improvement was also experienced in Medium-scale and the Large-scale problem with 4% $\cong$ \$ 129835.5 and 2% $\cong$ \$ 130522 , respectively.

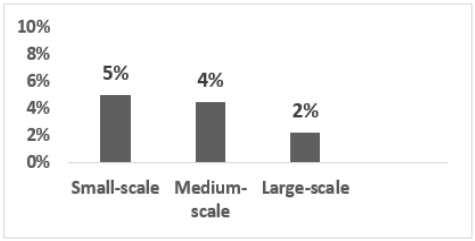
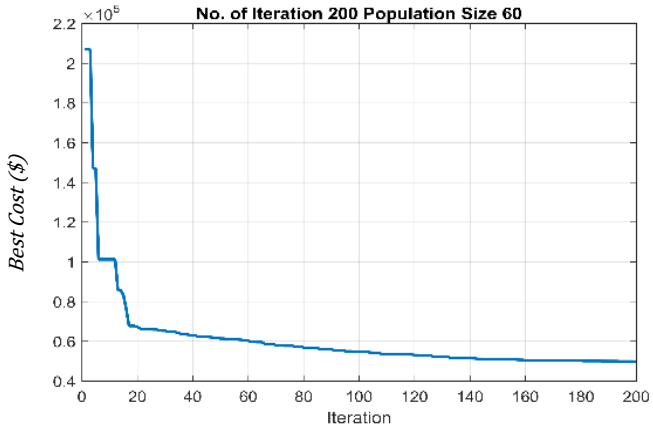
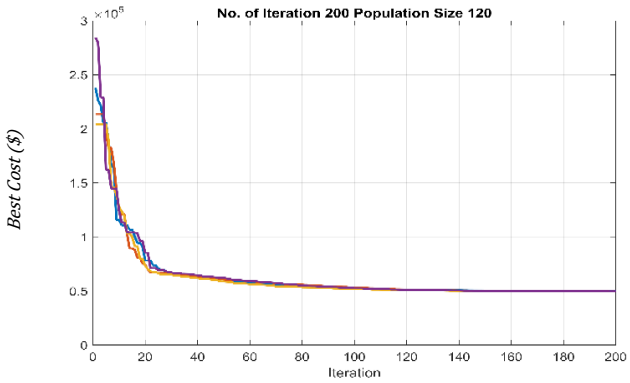


Figure 5 Improvement rates obtained from the tuning procedure

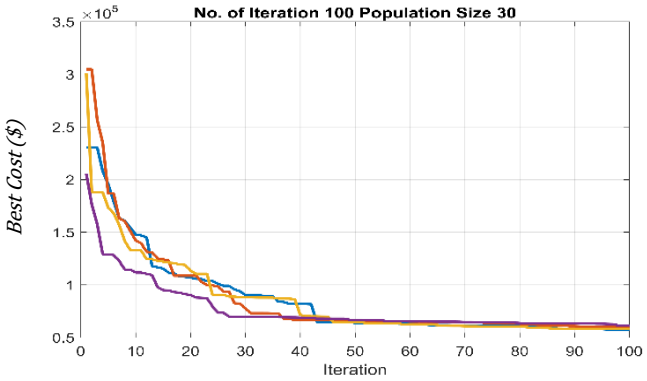
Also, the results obtained from the proposed SOA algorithm, and the GA solver associated with P_S, P_M, and P_L case studies are depicted in Figure 6-Figure 8.



(a) Proposed *SOA*

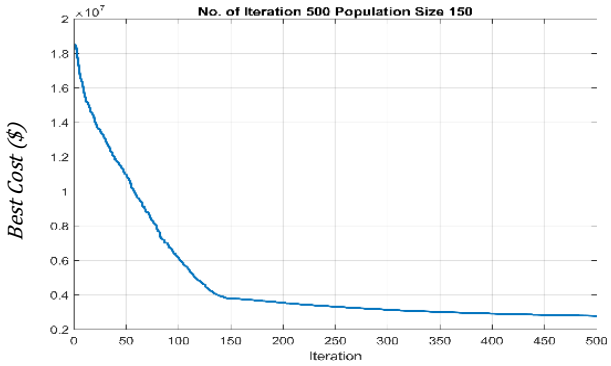


(b) Configuration 22

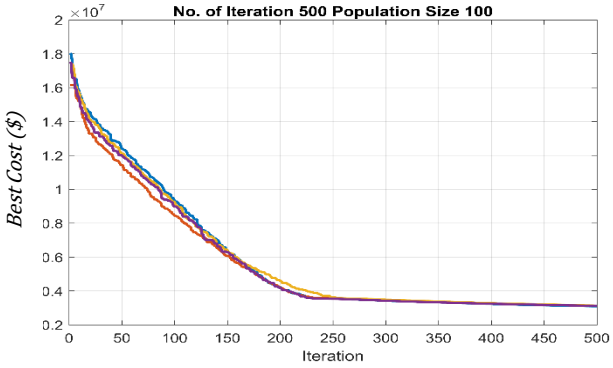


(c) Configuration 15

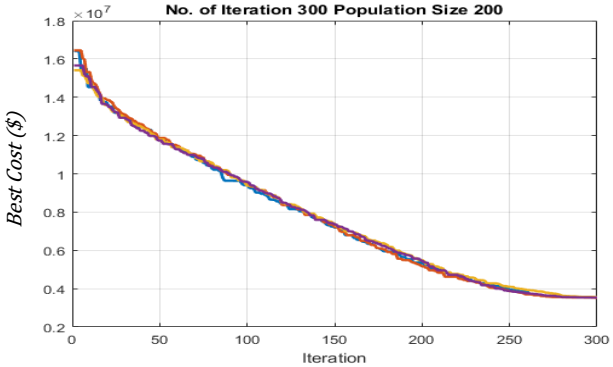
Figure 6 Results obtained from (a) the proposed SOA methodology, (b) the GA optimiser ($S22$) and (c) the GA optimiser ($S15$) for P_S



(a) Proposed *SOA*

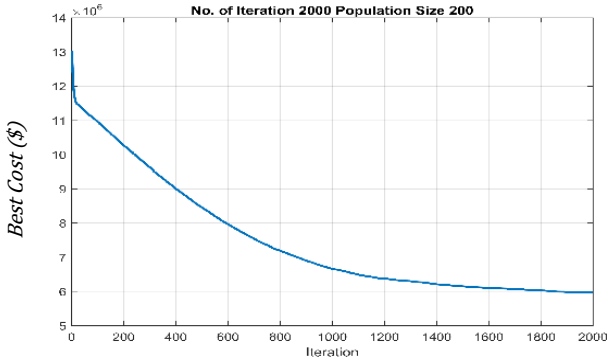


(b) Configuration 21

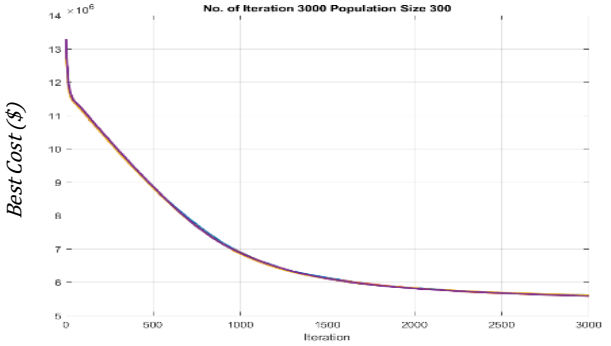


(c) Configuration 1

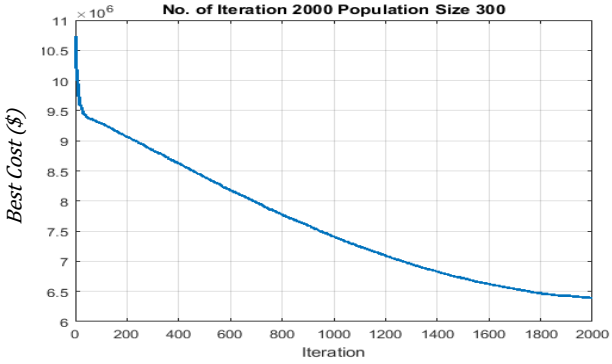
Figure 7 Total cost achieved from implementing (a) the proposed SOA methodology, (b) the GA optimiser ($S21$) and (c) the GA optimiser ($S1$) for P_M



(a) Proposed SOA



(b) Configuration 23



(c) Configuration 9

Figure 8 Total cost achieved from implementing (a) the proposed SOA, (b) the GA optimiser ($S23$) and (c) the GA optimiser ($S9$) for P_L

Table 18-Table 20 present the optimum quantities associated with each product family to be manufactured for consumers over the given planning horizon.

Table 18. The Optimum Solution for Small-scale problem

	P_1	P_2	P_3	P_4	P_5
T1	11	1	54	4	1
T2	10	11	1	5	80
Total	21	12	55	9	81

Table 19. The Optimum Solution for Medium-scale problem

	P_1	P_2	P_3	P_4	P_5	P_6
T1	136	314	362	220	450	276
T2	391	396	292	575	403	197
T3	369	658	557	574	464	349
T4	499	656	831	433	404	509
T5	577	622	727	681	1013	1086
Total	1972	2646	2769	2483	2734	2417

Table 20. The Optimum Solution for large-scale problem

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
T1	802	706	543	477	471	488	1026	768	670	590
T2	579	480	740	915	561	771	994	820	775	822
T3	710	811	917	608	877	703	952	791	946	1077
T4	1354	1128	630	1161	1058	1222	1090	1099	1427	1187
T5	1507	1771	1624	1429	1524	1229	1145	1537	1254	1554
T6	1685	1935	1762	1952	2055	1802	1848	1903	1397	1698
T7	1997	2192	2097	2037	2118	2435	1883	1918	2276	2854
T8	1904	2411	2159	2765	2271	2542	2309	2604	2437	1998
Total	10252	10371	10455	10606	10575	9608	9815	10685	10187	9990

8. Conclusion and outlook to future

In this paper, an advanced decision-making system for a class of CSN problems was proposed. A novel SOA algorithm incorporating GA as its optimisation module was designed for MPMPCSN problem. The robustness and effectiveness of the proposed methodology was verified through performing twenty-seven computational trials on three practical test problems at different granularity levels (small-scale, medium-scale, large-scale). In addition, a tuning mechanism was recommended to improve the quality of the obtained solutions that was affected by controllable parameters of the optimisation module. To this end, two statistical techniques known as ANOVA and Taguchi methods were utilised. The optimum levels associated to the controllable parameters of GA were determined as following: for $\boldsymbol{P_S}$, level 1(Roulette Wheel selection), level 1 (90% crossover), level 1 (10% mutation), level 2 (120 chromosomes) and level 1 (200 iterations), respectively; for $\boldsymbol{P_M}$ level 2 (Tournament selection), level 2 (85% crossover), level 1 (10% mutation), level 1 (200 chromosomes) and level 1 (500 iterations), respectively; For $\boldsymbol{P_L}$ level 2 (Tournament selection), level 1 (90% crossover), level 1 (10% mutation), level 3 (300 chromosomes) and level 1 (3500 iterations), respectively. The proposed SOA was resulted in 5%, 4% and 2% improvement in total cost of CSN associated to P_S, P_M and P_L problems respectively, in contrast to only using GA solver. Also, it was observed that the computational cost and time were reduced significantly.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are grateful to the Australian Mathematical Society (*Aust MS*) for providing the Lift-up fellowship which financially supported this work.

References

[1] Z. Hajiabolhasani, R. Marian, and J. Boland, "Consumer Supply Network Planning: Literature Review And Analysis," Journal of Multidisciplinary Engineering Science Studies, vol. 3, no. 3, pp. 1519-1538, 2017.

[2] Z. H. Abolhasani, R. Marian, and J. Boland, "Simulation-Optimisation of Multi-Product, Multi- Period Consumer Supply Network using Genetic Algorithms," in Intelligent Systems Conference (INTELLISYS), London,UK, 2017, pp. 34-44: IEEE, 2017.

[3] X.-S. Yang, S. Koziel, and L. Leifsson, "Computational Optimization, Modelling and Simulation: Past, Present and Future," in ICCS 2014. 14th International Conference on Computational Science, 2014, vol. 29, pp. 754–758: Procedia Computer Science.

[4] T. Hennies, T. Reggelin, J. Tolujew, and P.-A. Piccut, "Mesoscopic supply chain simulation," Journal of Computational Science, vol. 5, no. 3, pp. 463-470, 2014.

[5] R. A. Alive, B. Fazlohhahi, B. G. Guirimov, and R. R. Aliev, "Fuzzy-genetic approach to aggregate production–distribution planning in supply chain management," Information Sciences, vol. 177, pp. 4241-4255, 2007.

[6] N. Mustafee, K. Katsaliaki, and S. J. E. Taylor, "A review of literature in distributed supply chain simulation," presented at the Simulation Conference (WSC), 2014 Winter, 7-10 Dec. 2014, 2014.
[7] T. M. Pinho, J. P. Coelho, A. P. Moreira, and J. Boaventura-Cunha, "Modelling a biomass supply chain through discrete-event simulation∗∗This work was supported by the FCT - Fundação para a Ciência e Tecnologia through the PhD Studentship SFRH/BD/98032/2013, program POPH - Programa Operacional Potencial Humano and FSE - Fundo Social Europeu," IFAC-PapersOnLine, vol. 49, no. 2, pp. 84-89, 2016/01/01/ 2016.
[8] F. Campuzano and J. Mula, Supply chain simulation (A system dynamics approach for improving performance). Springer, 2011.
[9] G. Dellino, J. P. C. Kleijnen, and C. Meloni, "Robust optimization in simulation: Taguchi and Response Surface Methodology," Int. J. Production Economics, vol. 125, pp. 52-59, 2010.
[10] A. Huerta-Barrientos, M. Elizondo-Cortés, and I. F. d. l. Mota, "Analysis of scientific collaboration patterns in the co-authorship network of Simulation–Optimization of supply chains," Simulation Modelling Practice and Theory, vol. 46, pp. 135-148, 2014.
[11] X. Wan, J. F. Pekny, and G. V. Reklaitis, "Simulation-based optimization with surrogate models—Application to supply chain management," Computers and Chemical Engineering, vol. 29, pp. 1317–1328, 2005.
[12] J. Y. Jung, G. Blaua, J. F. Pekny, G. V. Reklaitis, and D. Eversdykb, "A simulation based optimization approach to supply chain management under demand uncertainty," Computers and Chemical Engineering, vol. 28, pp. 2087–2106, 2004.
[13] M. C. Fu, "Optimization via simulation: A review," Annals of Operations Research vol. 53, pp. 199-247, 1994.
[14] J.-H. Kang and Y.-D. Kim, "Inventory control in a two-level supply chain with risk pooling effect," International Journal of Production Economics, vol. 135, no. 1, pp. 116-124, 2012.
[15] S. M. Mousavi, A. Bahreininejad, S. N. Musa, and F. Yusof, "A modified particle swarm optimization for solving the integrated location and inventory control problems in a two-echelon supply chain network," Intell Manuf, 2014.
[16] F. T. S. Chan and A. Prakash, "Inventory management in a lateral collaborative manufacturing supply chain: a simulation study," International Journal of Production Research, vol. 50, no. 16, p. 15, 15 August 2012 2012.
[17] O. Labarthe, B. Espinasse, A. Ferrarini, and B. Montreuil, "Toward a methodological framework for agent-based modeling and simulation of supply chains in a mass customization context," Simulation Modelling Practice and Theory, vol. 15, no. 2, pp. 113-136, 2007.
[18] F. Longo and G. Mirabelli, "An advanced supply chain management tool based on modeling and simulation," Computers & Industrial Engineering, vol. 54, no. 3, pp. 570-588, 2008.
[19] A. Alrabghi and A. Tiwari, "State of the art in simulation-based optimisation for maintenance systems," Computers & Industrial Engineering, vol. 82, pp. 167-182, 4// 2015.
[20] P. Ghamisi and J. A. Benediktsson, "Feature selection based on hybridization of genetic algorithm and particle swarm optimization," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 2, pp. 309-313, 2015.
[21] A. E. Eiben, R. Hinterding, and Z. Michalewicz, "Parameter control in evolutionary algorithms," IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, vol. 3, no. 2, pp. 124-141, 1999.
[22] J. Sadeghi, S. M. Mousavi, S. T. A. Niaki, and S. Sadeghi, "Optimizing a multi-vendor multi-retailer vendor managed inventory problem: Two tuned meta-heuristic algorithms," Knowledge-Based Systems, vol. 50, pp. 159-170, 2013.
[23] H. Ding, L. Benyoucef, and X. Xie, "Stochastic multi-objective production-distribution network design using simulation-based optimization," International Journal of Production Research, vol. 47, no. 2, pp. 479-505, 2009.
[24] A. D. Yimer and K. Demirli, "A genetic approach to two-phase optimization of dynamic supply chain scheduling," Computers & Industrial Engineering, vol. 58, no. 3, pp. 411-422, 4// 2010.
[25] A. Nikolopoulou and M. G. Ierapetritou, "Hybrid simulation based optimization approach for supply chain management," Computers & Chemical Engineering, vol. 47, pp. 183-193, 12/20/ 2012.
[26] M. Seifbarghy, M. M. Kalani, and M. Hemmati, "A discrete particle swarm optimization algorithm with local search for a production-based two-echelon single-vendor multiple-buyer supply chain," Journal of Industrial Engineering International, journal article vol. 12, no. 1, pp. 29-43, 2016.
[27] E. M. Frazzon, A. Albrecht, M. Pires, E. Israel, M. Kück, and M. Freitag, "Hybrid approach for the integrated scheduling of production and transport processes along supply chains," International Journal of Production Research, vol. 56, 2018.
[28] J. Huang and J. Song, "Optimal inventory control with sequential online auction in agriculture supply chain: an agentbased simulation optimisation approach," International Journal of Production Research, vol. 56, no. 6.
[29] S. K. Shukla, M. K. Tiwari, H.-D. Wan, and R. Shankar, "Optimization of the supply chain network: Simulation, Taguchi, and psychoclonal algorithm embedded approach," Computers & Industrial Engineering, vol. 58, no. 1, pp. 29-39, 2// 2010.
[30] B. Unhelkar, Practical object oriented design. Thomson Social Science Press, 2005.
[31] J. Arthur F. Veinott, "Lectures in Supply-Chain Optimization," S. U. Department of Management Science and Engineering, Ed., ed. Stanford, California, 2005.
[32] B. Fahimnia, L. Luong, and R. Marian, "Genetic algorithm optimisation of an integrated aggregate production–distribution plan in supply chainsn," International Journal of Production Research, vol. 50, no. 1, pp. 81-96.
[33] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. MA: Addison-Wesley, 1989.
[34] N. M. Razali and J. Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP," presented at the Proceedings of the World Congress on Engineering 2011 Vol II, London, U.K., 2011.
[35] B. Fahimnia, "An Integrated Methodology for the Optimisation of Aggregate Production-Distribution Plan in Supply Chains," Doctor of Philosophy, Mechanical and Manufacturing Engineering, University of South Australia, 2010.
[36] R. Marian, L. Luong, and K. Abhary, "Assembly sequence planning and optimisation using genetic algorithms: part I. Automatic generation of feasible assembly sequences," Applied Soft Computing vol. 2, no. 3, pp. 223-253.
[37] J.-F. Cordeau, M. Gendreau, A. Hertz, G. Laporte, and J.-S. Sormany, "New heuristics for the vehicle routing problem," in Logistic Systems: Desing and Optimization, A. Langevin and D. Riopel, Eds. United States of America: Springer, 2005, pp. 279-298.
[38] R. M. MARIAN, "Optimisation of assembly sequences using genetic algorithms," DOCTOR OF PHILOSOPHY, School of Advanced Manufacturing and Mechanical Engineering, UNIVERSITY OF SOUTH AUSTRALIA, 2003.
[39] F. T. Chan, S. Chung, and S. Wadhwa, "A hybrid genetic algorithm for production and distribution," Omega, vol. 33, no. 4, pp. 345-355, 2005.
[40] Z. H. Abolhasani, R. M. Marian, and L. Luong, "Optimization of Multi-Commodities Consumer Supply Chain- Part 1- Modelling," Journal of Computer Science, vol. 9, no. 12, p. 16, 2013.
[41] H. Xing, X. Liu, X. Jin, L. Bai, and Y. Ji, "A multi-granularity evolution based Quantum Genetic Algorithm for QoS multicast routing problem in WDM networks," Computer Communications, vol. 32, pp. 386-393, 2009.
[42] C. A. C. Coello, G. B. Lamont, and D. A. V. Veldhuizen, D. E. Goldberg and J. R. Koza, Eds. Evolutionary algorithms for solving multi-objective problems, 2nd ed. (Genetic and Evolutionary Computation). New York: Springer, 2007.
[43] Jeong Hee Hong, K.-M. Seo, and T. G. Kim, "Simulation-based optimization for design parameter exploration in hybrid system: a defense system example," Simulation: Transactions of the Society for Modeling and Simulation International, vol. 89, no. 3, pp. 362-380, 2013.
[44] P. Genin, S. Lamouri, and A. Thomas, "Multi-facilities tactical planning robustness with experimental design," Production Planning & Control, vol. 19, no. 2, pp. 171-182, 2008/03/01 2008.

Appendix A

Table 21. The volume of product family P ($v_{P=1:I}^{G=\{M,L\}}$) used in Medium and Large -scale problem (non-linear Constraint)

Product Family	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
Purchase Cost ($)	72	51	16	74	100					
$v^{G=S}{}_i$	3	4	2	5	1					
Purchase Cost ($)	161	138	148	185	162	113				
$v^{G=M}{}_i$	5	6	1	3	5	3				
Purchase Cost ($)	103	167	159	197	171	160	118	109	178	104
$v^{G=L}{}_i$	1	2	3	3	4	5	4	1	6	2

Note: The volume of product family P_1 in the Medium-scale problem is $v_1^M = 5$

Table 22. The DEMAND QUANTITY ASSOCIATED TO product FAMILY i ordered by consumer j at PERIOD t for P_S

		RE_1	RE_2			RE_1	RE_2
T_1	P_1	50	60	T_2	P_1	64	74
	P_2	1	64		P_2	42	96
	P_3	32	74		P_3	18	29
	P_4	7	25		P_4	40	45
	P_5	10	78		P_5	24	69

Table 23. The DEMAND QUANTITY ASSOCIATED TO product FAMILY i ordered by consumer j at PERIOD t for P_M

		RE_1	RE_2	RE_3	RE_4	RE_5	RE_6	RE_7	RE_8	RE_9	RE_{10}	RE_{11}
T_1	P_1	58	4	37	98	76	81	8	4	52	94	76
	P_2	17	6	54	8	100	43	60	95	100	48	100
	P_3	15	81	72	59	19	73	92	77	86	24	97
	P_4	48	46	88	42	79	50	20	56	97	40	54
	P_5	91	39	33	31	20	81	44	19	68	71	97
	P_6	56	79	66	27	100	36	75	50	41	56	12
T_2	P_1	6	44	48	20	18	88	34	96	10	4	36
	P_2	31	55	26	20	97	79	60	55	47	21	79
	P_3	59	72	37	33	41	47	91	55	1	46	44
	P_4	54	2	67	89	85	82	71	32	92	13	44
	P_5	91	81	17	48	62	90	38	8	65	1	5
	P_6	55	15	28	41	38	43	74	19	1	73	5
T_3	P_1	10	49	47	44	33	3	37	86	69	35	81
	P_2	60	23	64	89	81	61	21	5	91	42	7
	P_3	25	23	92	40	100	12	45	70	62	16	96
	P_4	85	54	17	18	99	41	96	98	90	82	50
	P_5	86	77	72	64	13	89	13	29	20	63	76
	P_6	97	35	58	63	24	55	48	14	76	74	75
T_4	P_1	84	90	41	26	17	75	53	69	76	61	55
	P_2	16	59	4	33	19	70	33	24	99	86	21
	P_3	46	59	75	41	10	83	84	46	24	99	22
	P_4	62	86	16	41	33	83	82	39	53	93	33
	P_5	94	4	15	39	77	30	56	54	6	41	10
	P_6	84	89	61	61	24	31	27	100	76	1	75
T_5	P_1	75	90	75	79	23	17	5	69	81	80	30
	P_2	55	36	13	37	36	84	22	97	24	33	41
	P_3	34	55	83	75	29	17	40	44	94	23	87
	P_4	84	35	3	90	93	51	34	95	77	32	62
	P_5	56	63	42	25	6	100	23	1	83	59	100
	P_6	96	80	74	13	60	36	94	62	58	83	21

Table 24. The DEMAND QUANTITY ASSOCITED TO product FAMILY i ordered by consumer j at PERIOD t for P_i.

		RE_1	RE_2	RE_3	RE_4	RE_5	RE_6	RE_7	RE_8	RE_9	RE_{10}	RE_{11}	RE_{12}	R_{13}	RE_{14}	RE_{15}	RE_{16}	RE_{17}	RE_{18}	RE_{19}	RE_{20}	RE_{21}	RE_{22}	RE_{23}	RE_{24}	RE_{25}
T_1	P_1	32	92	33	68	98	7	10	39	73	48	33	38	69	4	3	27	67	22	26	41	64	59	18	70	61
	P_2	60	43	24	58	65	18	76	82	17	76	66	1	45	97	37	43	5	6	43	30	56	3	43	9	14
	P_3	45	40	48	40	85	63	87	52	65	99	2	5	60	10	20	23	12	64	77	46	47	44	94	20	51
	P_4	41	61	49	73	71	77	41	32	8	25	96	3	21	4	78	62	53	7	32	37	39	70	83	88	52
	P_5	59	18	20	38	12	54	99	59	24	21	17	91	78	49	22	19	94	23	60	24	83	90	51	38	11
	P_6	85	7	23	66	81	46	2	90	89	55	28	5	35	82	27	80	39	36	82	60	94	100	72	14	27
	P_7	29	38	71	27	31	78	77	38	92	69	22	23	52	74	80	66	79	45	44	76	71	4	87	40	7
	P_8	66	38	25	84	25	41	43	3	46	50	33	80	68	60	2	99	63	92	95	33	3	71	2	23	64
	P_9	17	5	53	77	60	41	62	44	67	74	29	91	60	31	76	8	79	34	7	31	58	96	97	88	31
	P_{10}	89	63	63	27	44	9	43	5	34	84	61	31	97	75	66	2	88	13	86	52	1	5	74	6	66
T_2	P_1	23	74	81	32	73	63	90	46	80	94	68	94	95	7	60	94	12	54	48	96	17	23	25	4	66
	P_2	18	30	87	13	26	49	96	86	91	93	66	40	45	28	75	83	17	7	47	6	27	63	61	88	79
	P_3	61	66	66	99	86	15	45	50	24	41	33	37	55	68	41	91	57	29	54	52	76	9	57	33	75
	P_4	85	45	16	40	44	23	16	42	70	16	76	99	59	92	12	7	34	37	58	98	60	71	55	85	26
	P_5	26	46	61	12	18	18	21	27	93	20	79	60	91	60	11	75	7	7	42	64	63	80	96	92	79
	P_6	17	62	74	10	91	38	43	62	69	67	52	34	50	34	32	58	21	35	61	37	34	10	67	4	77
	P_7	27	79	6	82	52	91	84	38	97	1	60	26	87	40	50	70	86	91	89	56	33	94	98	7	47
	P_8	18	41	62	54	35	31	84	44	43	67	6	41	41	98	65	23	76	10	58	85	94	15	33	60	65
	P_9	53	74	6	25	14	2	31	46	35	43	84	20	89	17	99	52	78	57	50	72	17	44	79	13	78
	P_{10}	62	87	81	26	62	3	11	80	16	32	46	21	41	51	76	12	19	84	16	95	78	6	83	76	59
T_3	P_1	38	50	78	66	43	80	19	41	85	53	42	80	86	11	38	30	96	48	94	97	80	41	57	91	43
	P_2	32	73	63	53	59	33	88	6	11	75	10	6	15	16	86	6	62	84	41	2	20	67	59	57	17
	P_3	9	28	34	33	77	74	74	60	48	40	10	76	34	82	12	25	14	68	48	6	45	3	61	65	61
	P_4	65	64	88	47	27	26	99	66	62	93	93	24	71	33	12	95	26	28	43	45	71	25	80	97	64
	P_5	1	89	71	52	35	48	96	96	4	35	46	18	54	4	54	60	2	27	99	14	99	29	70	78	87
	P_6	43	73	37	77	60	2	73	5	6	20	82	89	40	27	22	56	82	98	80	7	42	28	57	47	88
	P_7	5	33	46	57	58	9	68	36	81	82	61	50	18	58	78	32	2	32	39	16	53	9	18	15	96
	P_8	29	42	72	71	32	92	87	6	64	10	32	87	72	81	100	12	19	28	77	4	67	20	55	44	86
	P_9	99	67	73	22	97	89	20	93	29	30	39	23	17	84	47	63	82	97	58	17	96	45	21	67	89
	P_{10}	48	95	41	50	67	41	12	49	72	80	39	50	92	50	49	3	25	99	66	87	62	84	55	10	44
T_4	P_1	81	34	10	12	36	99	33	44	95	86	71	58	42	28	9	64	54	1	52	99	32	19	26	62	90
	P_2	82	61	59	94	46	54	33	75	55	64	23	74	77	100	15	9	96	20	45	91	85	45	27	86	16
	P_3	44	80	25	49	87	96	71	48	49	95	59	10	80	35	61	81	33	57	11	28	80	21	31	49	95
	P_4	48	46	75	32	13	62	70	7	98	25	44	43	59	38	89	45	32	71	76	25	22	75	24	15	26
	P_5	33	16	90	51	12	2	79	88	88	46	14	46	24	11	37	42	5	78	6	89	95	32	74	19	27
	P_6	22	37	45	26	71	32	1	16	80	10	23	54	83	85	91	17	71	83	17	93	10	10	88	4	88
	P_7	17	52	39	79	6	82	18	100	23	92	62	90	43	11	81	88	55	23	38	40	43	33	69	89	50
	P_8	73	28	75	56	53	37	69	92	79	20	54	92	7	87	4	54	74	60	66	78	100	100	82	15	76
	P_9	28	58	71	75	28	27	22	52	59	57	51	74	94	26	28	86	29	18	99	97	70	9	58	51	22
	P_{10}	60	52	2	45	59	18	88	11	35	28	51	50	85	60	88	40	43	25	6	79	37	53	51	64	9
T_5	P_1	47	12	3	29	57	31	59	71	80	37	70	12	87	25	72	70	70	84	47	85	12	49	60	32	17
	P_2	75	16	41	37	25	41	81	98	41	73	38	28	24	52	45	61	51	54	63	98	27	31	46	76	19
	P_3	31	44	98	4	66	2	14	76	16	86	36	47	66	21	77	46	38	10	57	45	26	92	71	58	53
	P_4	64	89	8	71	84	2	47	99	90	65	19	73	28	98	62	59	38	7	43	7	81	9	53	31	29
	P_5	57	86	81	91	15	71	44	91	6	80	38	34	63	12	28	57	94	20	95	44	26	79	82	37	89
	P_6	33	58	9	41	72	94	86	29	21	6	38	47	14	21	89	46	24	33	80	75	86	75	64	40	16
	P_7	5	56	21	11	98	42	29	19	96	33	44	9	45	58	49	83	29	21	46	49	88	58	15	35	20
	P_8	43	88	21	48	83	41	11	50	70	28	8	12	93	97	82	39	19	62	19	39	47	32	38	79	52
	P_9	89	53	51	50	76	77	53	9	29	36	26	20	74	3	49	58	27	2	5	43	56	62	87	50	65
	P_{10}	88	5	34	34	2	51	34	13	6	64	8	68	26	27	24	4	46	36	2	86	87	70	41	58	74
T_6	P_1	29	71	70	31	35	30	44	2	57	2	68	100	87	30	8	19	41	72	53	68	35	17	59	12	19
	P_2	63	84	72	4	94	70	46	15	57	66	53	12	44	92	32	16	70	9	44	85	2	44	97	93	90
	P_3	27	52	34	60	66	53	25	69	42	48	33	87	5	31	48	83	47	47	59	38	84	67	17	73	91
	P_4	5	41	92	97	82	83	93	83	93	83	29	86	36	70	98	35	81	67	11	93	67	59	29	7	82
	P_5	5	46	99	88	53	99	32	12	17	69	100	7	6	92	80	81	90	44	55	31	56	46	57	49	94
	P_6	87	2	89	80	73	77	48	37	78	62	10	77	69	19	14	71	62	10	85	14	25	23	39	60	56
	P_7	85	9	91	96	83	52	97	37	98	1	61	28	99	94	45	99	94	38	77	51	33	6	87	27	78
	P_8	6	71	70	62	85	47	33	42	48	69	40	81	55	3	65	89	12	82	100	81	84	37	93	14	87
	P_9	81	7	56	100	6	85	22	3	86	41	9	84	74	27	2	26	87	55	22	20	96	98	17	88	10
	P_{10}	75	91	24	84	87	22	7	60	93	5	90	46	13	80	51	11	3	9	22	77	76	74	39	94	69
T_7	P_1	25	91	72	2	28	5	42	15	76	59	32	16	45	82	62	5	87	79	6	25	5	36	19	35	98
	P_2	30	51	67	86	37	83	69	12	73	58	77	97	2	57	31	63	7	46	94	71	21	80	99	2	42
	P_3	28	80	5	30	95	33	83	3	50	64	83	69	66	40	96	71	90	31	17	95	64	67	59	16	23
	P_4	33	68	59	88	17	33	97	18	6	9	47	50	85	54	48	65	16	88	60	92	73	34	49	80	76
	P_5	97	45	64	18	57	85	92	66	15	70	8	16	44	76	53	48	37	61	49	73	3	44	13	84	3
	P_6	8	16	72	62	96	2	79	49	23	83	49	89	86	29	23	61	33	100	42	15	75	73	40	89	1
	P_7	69	63	76	47	60	23	17	48	81	34	42	76	34	47	17	11	69	55	79	37	5	8	28	80	81
	P_8	81	98	24	72	29	29	8	30	71	87	31	64	72	39	67	70	95	6	99	47	48	92	23	96	25
	P_9	34	14	10	31	52	86	35	74	12	8	4	19	28	32	73	17	2	90	36	72	16	63	1	96	36
	P_{10}	74	5	60	38	33	65	68	27	14	8	75	9	72	16	24	54	44	69	100	75	1	15	69	68	59
T_8	P_1	65	25	64	75	12	70	77	47	81	4	54	21	90	67	71	22	62	27	19	46	53	39	46	62	21
	P_2	39	33	44	54	23	92	43	12	3	71	69	86	69	61	67	10	9	10	32	24	99	21	48	31	44
	P_3	24	30	22	20	75	71	51	78	72	64	27	5	15	15	66	30	78	36	100	64	45	13	68	62	16
	P_4	44	74	2	68	100	47	90	62	34	45	72	52	5	51	14	10	6	56	100	30	68	28	93	3	30
	P_5	52	48	97	50	51	35	43	7	82	53	75	18	94	37	81	98	21	69	11	63	22	55	47	70	14
	P_6	40	57	19	61	64	54	93	6	84	33	48	29	58	25	100	4	23	95	6	67	11	85	13	22	14
	P_7	79	52	91	16	70	2	38	96	13	54	50	90	84	15	97	87	66	25	93	31	23	99	80	5	25
	P_8	49	39	52	39	49	57	82	52	6	34	9	77	71	97	4	16	56	73	7	25	99	45	82	7	35
	P_9	30	18	42	47	92	10	36	58	86	45	35	24	76	7	78	79	65	76	30	69	84	35	33	34	24
	P_{10}	75	36	37	23	7	47	76	78	29	51	87	86	74	100	95	17	69	47	93	84	94	70	72	60	65

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 478-487 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Numerical Simulation of One Pavement Structure of Polyethylene Terephthalate Submitted to Static Point Loads

Juan Sebastián Gilart Pulido*, María Camila Zambrano Bello, Ender Jhobany Orduz Duarte

Program of Civil Engineering, Universidad Católica de Colombia, 111311, Colombia

ARTICLE INFO

Article history:
Received: 15 August, 2018
Accepted: 07 October, 2018
Online: 25 October, 2018

Keywords:
Polyethylene Terephthalate
Vertical displacements
Simulation
Structure analysis
Alternative pavement
Sustainable materials

ABSTRACT

Plastic is highly polluting nowadays and is found in many disposable elements used for people. Plastic materials are disposed inadequately and large quantities of them become pollution in the ground, water, and air, affecting plants and animals. Most of the plastic is Polyethylene Terephthalate (PET) which is used in many activities. Resistance, support capacity, and flexibility are some characteristics that make PET useful in road engineering. This research shows a numerical simulation of three kinds of structures for pavements in PET. The simulation uses 3D models that were analyzed in the software SAP2000. The material characteristics of the recycling PET were taken from industries and analyzed in the laboratory. The model used in the simulation was composed of vertical loads for traffic type A. Loads were applied in the border, 0.75 meters and 1.5 meters from the border. Three types of structures were analyzed with three, four and five supports respectively. Structure type 2 with four supports displayed the best results regarding vertical displacements in all the elements. Structure type 2 was simulated with six different thicknesses (0.06, 0.08, 0.10, 0.12, 0.14 and 0.16 meters). Optimization of this material and the resistance of the structure showed the 0.10 meters thickness as the best due to its admissible vertical displacements and required the lowest quantity of material for producing it.

1. Introduction

Transporting people and resources are two of the most important elements of economic development in society. Comfort, security, and speed are three parameters required for users of the roads. The concrete asphalt is one of the most popular materials used in pavements in order to have high quality and adequate traffic service nowadays [1]. Environmental damage caused by asphalt production and building roads in concrete asphalt is an important topic studied lately. Production, transport, and storage are activities that produce significant pollution. The Environmental Protection Agency of the United States has pleaded worldwide activities to underwrite production controls, environmental impact reduction, and more policies of environmental care around the world. Researches account more than 500000 kilograms of emissions to the atmosphere in pollution with volatile compositions, organic materials and so on [2]. Environmental damage is an important topic at the present time and society should find new strategies for controlling all different kinds of pollution in it. The traffic and the transport of resources and materials are activities that induce high quantities of pollution, especially into the air. Using synthetic materials and other kinds of additions for making the road`s concrete is one of many development possibilities to build road infrastructure around the world [3,4]. The addition of polymeric materials to concrete are implemented in order to reduce the volume of waste in landfills and to recycle different kinds of polymeric elements. Sustainable activities have served as a way of making new construction materials and resources for customers [5,6]. Reusing materials, used currently, could be another strategy to reduce humans' impact on the Earth [7].

Polymers are versatile materials and many products are made with it. However, polymers are also highly polluting materials because a large number of disposable elements are used daily and not all of them are recycled or reused in new activities. As a result, those elements are transported to the landfill and increase the contamination problems in land, water, and air; it can attack different species of animals in nature. Synthetic polymers are potentially polluting because their degradation time is so long and the oil is an irreplaceable resource. Also, the chemical composition and the high quantity of waste during the production is a considerable issue [8].

This research shows a strategy proposed for making pavement structures in recycled plastic. The Polyethylene Terephthalate is a material easy to find in the modern world. Some alternatives made in the past have identified materials used for different activities and their characteristics can offer functionality and quality service

*Corresponding Author: Juan Gilart, Email: jsgilart@gmail.com

in roads infrastructure as a pavement structure. The process shows three kinds of structures evaluated in the software SAP2000 with the finite element methodology. The simulated material is recycled plastic, a unique resource for the structure. The software seeks to evaluate the model of the structures with the application of static and point loads.

3. Synthetic polymers materials and their recycling

Gaelle Gourmelon research, "Global Plastic Production Rises, Recycling Lags," discusses the behavior of the polymeric industry around the world during the period 1950-2012. This industry produces more than six hundred thousand dollars each year. Transport, construction, health, education, and food industry are some of the sectors of the economy where plastic is highly used. For example, for making one automobile the industry uses 152.41 kilograms of plastic, while in 1960 it only used 9.07 kilograms. Of course, technology advances look for high quality and less cost: the automotive industry uses polymers in important parts of the cars. During 2013, the US automotive industry nearly used 300000 million kilograms of plastic materials in automobile parts [9].

During 2012, Europe recycled 26% of its polymeric materials. Additionally, during the same year, 36% of this material was incinerated. Around 38% of the polymeric materials were taken to landfills. Austria, Belgium, Sweden, and Switzerland governments regulate recycling, and citizens are required by law to recycle plastics and other materials. These countries have the highest rates for recycling plastic materials and most of those materials are transformed or incinerated [9].

Microplastics, little pieces of plastic in millimetric size, are highly polluting. These microplastics have become a major problem because they are scattered easily and are potentially dangerous for animals and the forest [10]. Around 275000 million kilograms of plastic waste were generated in 2010 in 192 countries with coasts. It is estimated that 12700 million kilograms of that garbage polluted the ocean [11]. Some estimations calculate around 268900 kilograms of garbage floats on the ocean's surface, without considering the material below the surface [12].

In Colombia, the DNP (by its acronym in Spanish Departamento Nacional de Planeación) estimates that 11600 million kilograms of garbage is produced annually and just 17% is recycled. Some policies are applied in order to reduce the impact of this waste on the environment. The DNP generates the Integral Policy for Management of Solid Waste with the horizon in 2030. This policy calls for the implementation of one official plan for the integral management of solid waste as a public policy, with interest for all the population, in order to foster an economic and sustainable development, transformation of the society and mitigation on the environmental impact in the country, and of course, for all the planet. Research shows that 2800 people work in Colombia in the recycling business, from the collection of pieces to fiber production in recycling factories [13]. Not only are they factory workers, where the principal resources are the recycled plastic, but most of them are not formal employees. This topic is predominantly studied in socio-economic research because it is an important aspect of a country's economic development. Furthermore, ground pollution is another environmental problem. Plastic materials degrade over a long period of time and cause ground pollution that affects the growth and development of animals and plants [14]. This prediction of a sanitary emergency in most of Colombia's cities by 2030 highlights that the majority of the waste in Colombia is disposed in landfills, and eventually, those landfills will not have capacity for any more garbage. Considering the DNP`s information in 2016, Colombia needs better waste management, to change its consumption of resources and to transform its industries as soon as possible, in order to promote a productive and sustainable economy.

4. Reused polyethylene Terephthalate as an absolute material for pavement structures

Roads are important for economic development: industries, tourism, commerce, agriculture, and so on, are important sectors in developing countries. High quality roads permit faster traffic, fewer transportation costs and less pollution from emissions into the air. Today, one of the most important elements for defining economic development is the investment in quality road construction, faster traffic and safer roads: the base for increasing productivity [15]. Synthetic polymers are used frequently for making reusable materials in construction. Masonry is one of the most important activities where plastic is used. Recent research has attempted to use bricks made with plastic in building construction in order to reduce the impact of plastic on Earth. Polyethylene Terephthalate is the most prevalent polymer to modify construction materials. Using plastic bricks in buildings curbs challenges in traditional construction methods in developing countries and reduces the environmental impact of garbage in landfills [16]. Polymers are also frequently added to an asphalt mix. The addition of polymers to asphalt concrete changes its characteristics, especially the mechanic behavior, durability and thermal susceptibility [17]. Likewise, polymers are used as concrete aggregates. This kind of mixture changes their resistance and durability. However, this material should be used carefully in big buildings because of changes in their composition [18].

Polyethylene Terephthalate has valuable qualities, especially its durability and resistance to abrasion, it's an adequate answer to chemical attack, resistance to weather conditions and can be used in construction. They show better conditions and higher potential use in the construction activities compared to with other materials [19]. Construction factories around the world use commonly plastics in many applications. New infrastructure projects use lighter materials made of polymers to construct architectural finishes and residential public service networks. Nevertheless, recycling is not a frequent activity in developing countries, despite it's easy management and material characteristics. This research shows a new possibility for recycled plastic to develop road projects in places of light traffic loads. Also, the research as a first approximation in one type of structure developed only in plastic. Using recycled plastic in high quantity helps society in two important ways: first, in development of new materials in pavements for new road projects; second, plastic pavements demand high bulk from plastic materials that reduce the quantity of waste in landfills.

5. Characteristics of the model

Different kinds of structures are used in road pavements. Asphalt concrete is the most used material in developing

countries. Also, rigid concrete made with cement is used frequently. Another different pavement called articulated pavement uses masonry on the surface of roads with light traffic. All of these kinds of structures are constructed by layers. In the bottom of all projects, subgrade material is made up the natural soil. Sometimes, low resistance materials are changed or modified with some additions. Over the subgrade are used arid material, produced by rock's trituration (natural or artificial), builded in layers according designs. The top layer normally is made of concrete or masonry. The research proposes to change classic pavement structures of asphalt concrete or hydraulic concrete to plastic elements over arid layers. The Figures 3, 4 and 5 show some possibilities of structures where the plastic is in the top and in the bottom, but inside the structure is empty and there are ducts that can be used as channels for water or electrical or hydraulic networks.

The model presented was developed in pieces with length and width of 3 meters, a total height of 0.5 meters, and a horizontal and vertical thickness of 0.14 meters. The proportions used were taken from base recommendations for rigid pavements that INVIAS (Instituto Nacional de Vías) proposes for these types of pavements [20].

The Colombian norm NTC 237 recommends tests in density and absorption in plastic materials as Polyethylene Terephthalate. This research tested recycled Polyethylene Terephthalate from plastic bottles. Its density was 1219 kilograms per cubic meter and the absorption 0.28%. Figure 1 shows the material tested in the laboratory at Universidad Católica de Colombia in 2017.

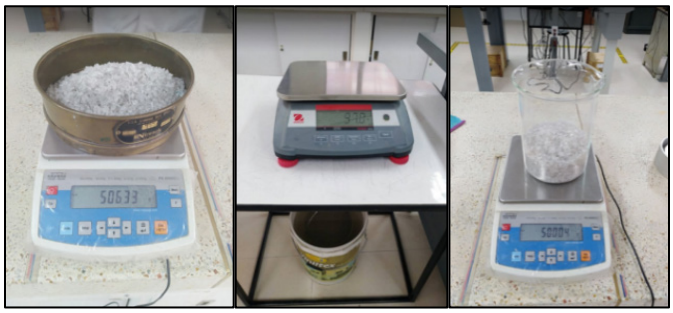


Figure 1. Polyethylene Terephthalate tested in the laboratory.

A preliminary model proposed recycled plastic for the entire structure as a first approximation through numerical simulation. Its specific weight was 1219 kilograms per cubic meter, elasticity 3.70^{8} Pascal, Poisson distribution 0.37 and thermal expansion 80^{-6} $Kelvin^{-1}$ [21]. Density and other characteristics were analyzed in the civil engineering laboratories of Universidad Católica de Colombia. Plastic material was tested in flexion in Test Three Points with a prototype in reused polyethylene terephthalate. The dimensions were: length 0.25 meters, width 0.05 meters and 0.01 meters thickness.

Prototype's flexion test was developed in an electronic universal machine and the graphics load-deflection were extracted from its software. Figure 2 shows the graphic load vs displacement of the material. Plastic behavior is in the linear part of the graffic between 0 meters to 0.03 meters deformation. The load applied for this deformation was 380 N.

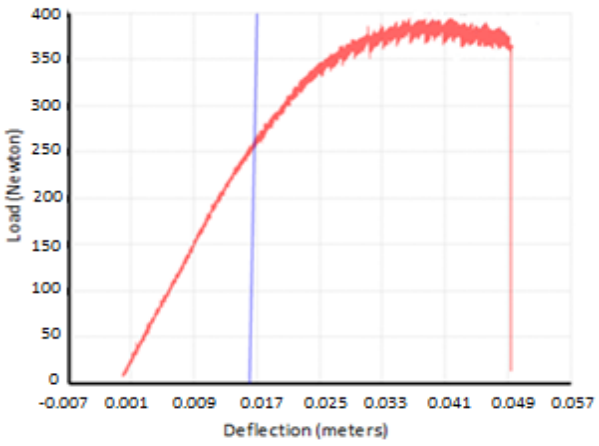


Figure 2. Load vs deflection in flexion test.

The laboratory results support the inference that recycled plastic can be used in the entire structure and the process presupposes the material offers resistance, flexibility, and functionality in structures of pavement.

5.1. Structure type 1.

Figure 3 shows the Structure type 1 model. This included three supports between the top and the bottom. Each support is 0.14 meters wide and as long as the element. Two of the supports are in the borders and the other is in in the middle.

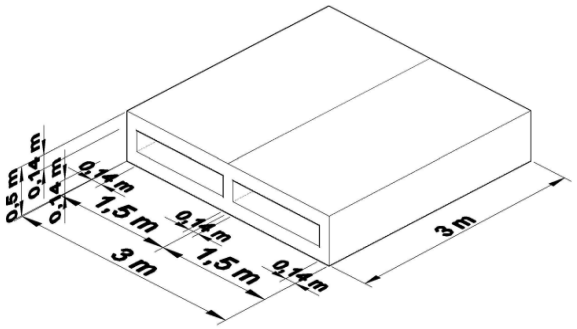


Figure 3. Structure type 1.

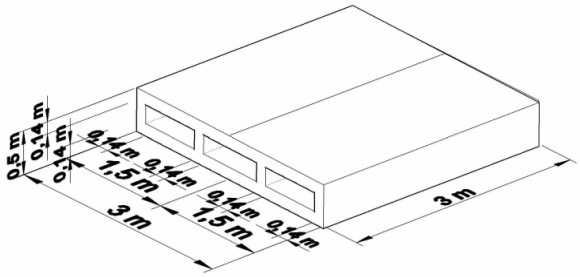


Figure 4. Structure type 2.

5.2. Structure type 2.

Figure 4 shows the Structure type 2 model. This included four supports between the top and the bottom elements. Each support is 0.14 meters wide and as long as the element. Two of the supports are in the borders and two are one-third (1.0 meters) to each border.

5.3. *Structure type 3.*

Figure 5 shows the Structure type 3 model. This included five supports between the element in the top and the bottom. Each support is 0.14 meters wide and as long as the element. Two of the supports are in the borders and the other one is in the middle. Also, two of the supports are in the quarters (0.75 meters) to the borders.

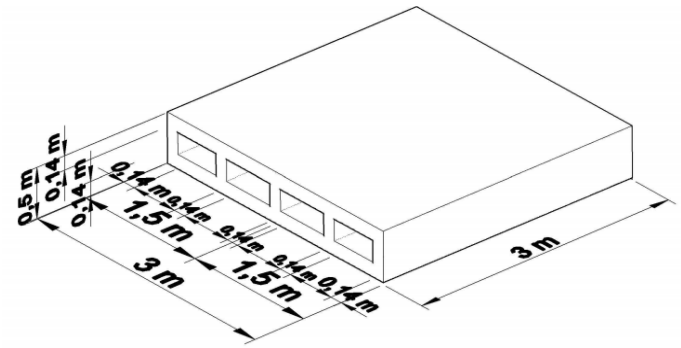


Figure 5. Structure type 3.

6. Simulation.

The process developed in this phase of the research was the simulation in the software SAP2000. This software uses the methodology of limit elements and with a sophisticated, intuitive and versatile tool used in design for engineering in different applications. This software enabled the creation of 3D elements for modeling analysis, including research parameters used to simulate one structure in plastic as real as possible. This is the first approximation for modeling structures without the arid layers of support. The subgrade was assumed to be a non-deformable material; it was an initial hypothesis for modeling, albeit the majority types of soils are deformed under loads. However, due it begin the first model, research analyzed only the element in plastic. The software permitted analysis of displacements in many points of the structure because the model can be divided in a high quantity into little cubic elements. Parameters are included in SAP2000 and the network is developed with 4 nodes.

Once the model is made, all elements can be divided into a network, and each node permits the identification of displacements under loads in different positions in the structure. This is achieved in multiple steps. First of all the element is divided into multiple series of elements. The series are determined by the required precision. The precision was connected by a discrete number of nodal points. Subsequently, a matrix is developed. The matrix relates variables in the nodes for each one of the divisions. Afterwards, the matrix is combined in a set of algebraic equations that describe the behavior of the structure analyzed. Finally, the set of algebraic equations is solved by a matricial method [22].

Elements from the general structure are drawn with shells. In the software, each node of the top shell has five grades of liberty with deformation. Three of them are translations (axes X, Y, and Z) and the other two are rotations (axes X and Y) [23].

Characteristics of shell elements showed the closest for the analysis, so the model was drawn in shells for this research. Analysis of deformations in Polyethylene Terephthalate with loads over the surface can be easily studied and it is possible in its study to measure reactions, deformations, and strengths in each node of the structure. Furthermore, elements were defined as "thin" based on the relation between thickness and length; the relation showed the length was over 20 times greater than the thickness. This relation is important to determine if elements work thin (more flexible) or thick (more rigid). This definition permits to SAP2000 the highest detail in the calculations developed in numerical simulation [24].

The element in the bottom was defined with restrictions of all displacements. For this simulation the soil is non-deformable and the structure does not suffer other effects. Using this assumption, though not real, provides the possibility for the first approximation of simulation. The target with this assumption is to analyze only the structure in recycled plastic and not other materials. Also, SAP2000 is a tool used in materials with different behavior as shown by the soil. Figure 6 shows the structures type 1, 2 and 3 modeled in SAP2000 with restrictions as supports in the bottom's element.

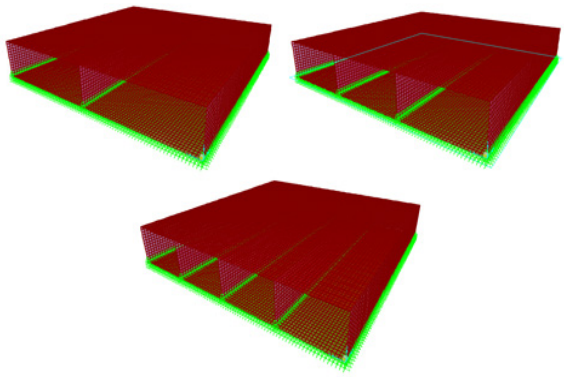


Figure 6. Simulation conditions of structures type 1, 2 and 3.

Loads applied in the simulation were defined in accordance with INVIAS to point load generated by trucks with a single axle and double tire. The norm defines 11000 kilograms for this axle divided into two point loads over the pavement with 58000 Newton each tire. This value is assumed even though it is superior to vehicles type A. The loads in real parameters could be present in tracks for passengers and light loads.

Loads applied simulate fires that produce point loads in each axle over top shell in the structure. Model in SAP2000 permits the location of loads in each node of the mesh. The structure was divided into nodes each 0.15 meters and one load per simulation. Using this mechanism shows the highest detail as possible, in order to identify the place where the structure could be more vulnerable. Furthermore, the load was applied in the border, in the middle (1.5 meters of the border) and each quarter (0.75 meters of the border) in the parallel direction of traffic as is shown in the Figure 7. Also, each load was simulated independently as is shown in Figures 8, 9 and 10 for structures type 1, 2 and 3. Each kind of structure was analyzed with the same parameters of the load in order to do the comparison of vertical displacements in the top for the three types of structures. Logically, it is possible to think that Structure type 3 would have fewer displacements. However, the intention with all simulations is to identify the

structure with permissible deformations using the minor quantity of plastic in its construction.

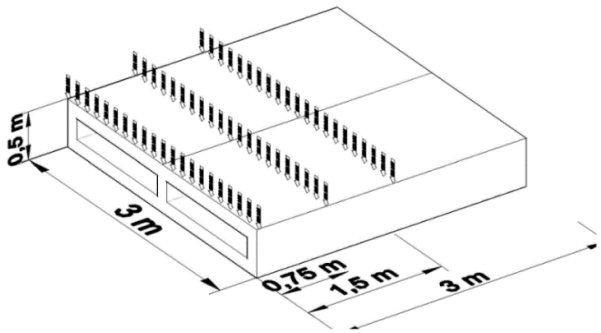


Figure 7. Isometric of application to loads on structure type 1.

Developing the simulations required 189 times with one load. The output information by SAP2000 was saved in data processor Excel and the relevant information was extracted. This information is selected to compare each node network.

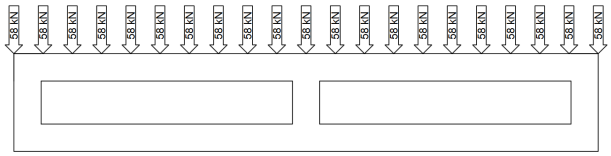


Figure 8. Loads application in cross-sectional view in structure type 1.

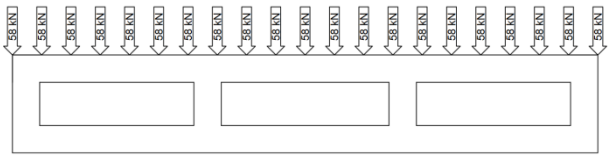


Figure 9. Loads application in cross-sectional view in structure type 2.

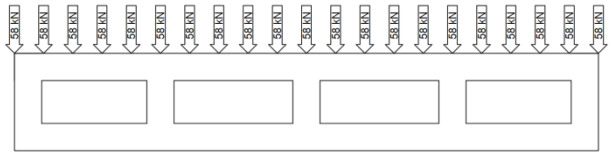


Figure 10. Loads application in cross-sectional view in structure type 3.

7. Results and analysis

Vertical displacements in the nodes are shown in a discrete network. The point load (58000 N) was applied over the top surface in the structure for each simulation. The information permits the comparison of behavior in all the points in the structure when the load is applied to different points over it. The purpose was to identify the critical point of the structure, or the place where the highest displacement is shown and defined by whether it is possible to develop a plastic structure in these conditions of simulation. Figure 11 shows an example of how the structure was analyzed, each simulation in two axes longitudinally and transversely.

Vertical displacements are typical behavior from plastic materials loaded with gravitational forces. The load perpendicular to the structure and deformations are shown for all of the structure. Some of the points display performance compression and others traction. Deformations are produced by a thermodynamic process. Energy is accumulated internally and is transformed in elastic potential energy. So, it is important to know the elasticity module and use it correctly in order to find high accuracy.

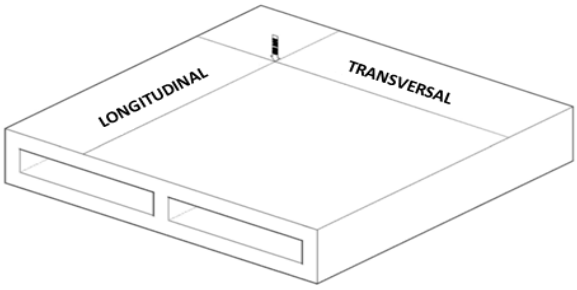


Figure 11. Axes for measuring results.

Figure 12 shows, without a real scale, how one load can affect entire structure. So, it is necessary to analyze all points and to compare displacements at the same point in the structure. This research shows displacements in the top. There are more vulnerable places where the structure could fail.

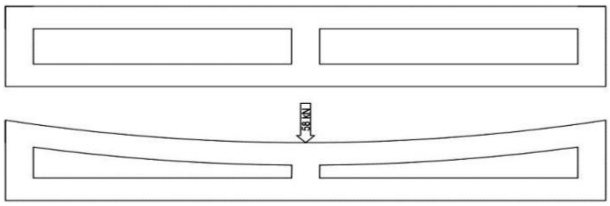


Figure 12. Model of displacements cross-sectional in structure type 1.

The same load reviewed in the direction parallel to the traffic, it has a different behavior in accordance with the point where is applied the load. Figure 13 shows one model without scale about how are the displacements in the structure when the load is applied to the border of the structure.

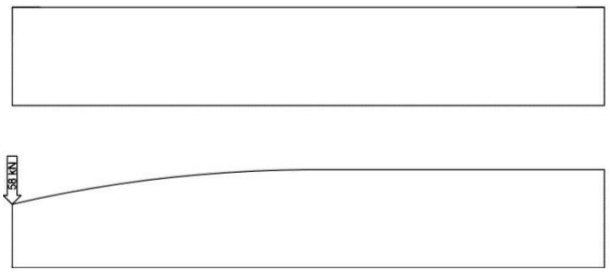


Figure 13. Model of displacements in the parallel direction to the traffic in structure type 1.

Models used were defined in three cases. Case A, when the loads are applied in the border. Case B with load on 0.75 meters of the border and Case C with loads on the middle of the structure (1.5 meters of the border). Vertical displacements in structure type 1 were taken in the three cases and compared in Figure 14. It is possible to identify that in the cases B and C, vertical displacements are shorter in comparison with case A. The results obey to the continuity of the material. For example, in the border (case A) deformations in the middle of the empty space are more than 0.0035 meters. While in the quarters (case B) or in the middle (case C), vertical displacements are around 0.00025 meters. In other words, vertical displacements are between 60% and 73%

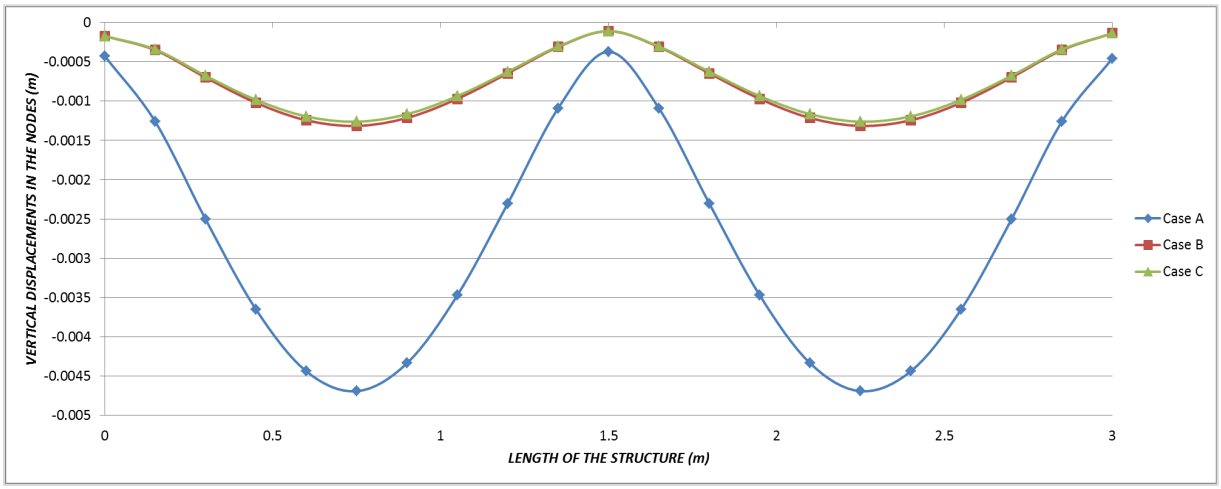
less in continue material. In this case, it is important to identify different mechanisms for doing a complete structure in order to transfer the loads adequately and get less deformation in the borders of the elements. It is possible to identify also that on the supports, vertical displacements are depreciable; the solid material shows the high resistance of contraction to continue material.

Consequently, when it analyzed the structure type 2, the behavior is completely similar to the first type of structure. Deformations on supports are very small and in the middle of the empty spaces are highest as is shown in Figure 15. In this case, three cases are analyzed as well. In the border (case A) in a quarter to 0.75 meters (Case B) and in the middle to 1.5 meters of the border (case C) all of them were studied by the same magnitude of load but the difference is the number of supports throughout the structure.

In the Figure 15 can be seen that vertical displacements are significantly smaller than structure type 1. In the border, the highest deformation was around 0.00168 meters. Also, the middle vertical displacements are around 0.00025 meters. The magnitudes permit to infer that this type of structure can have an adequate behavior for difficult conditions with important applications in roads engineering. Deformations conserve the same relation between 60% to 74% less in the quarter (case B) and the middle (case C) in comparison with the border's vertical displacements.

Clearly, it is easy to identify that where the material is homogeneous as under the supports, the vertical displacements are considerably minors than in the border where it showed the highest deformation. Furthermore, with this results in the border analyzed vertical displacements, it is possible to infer that the behavior is completely defined by the boundaries condition and especially when the material is discontinuous. Making a solid structure is a possibility for fewer deformations. However, it is important in this phase of the research to identify the behavior in all the structure because of the necessity of force's transference in completely necessary in roads engineering.

Structure type 3 was analyzed in the same way. The case A with loads throughout the border, punctual loads on 0.75 meters of the border (case B) and in the middle of the structure (case C). Figure 16 shows the envelope of vertical displacements in each case. The



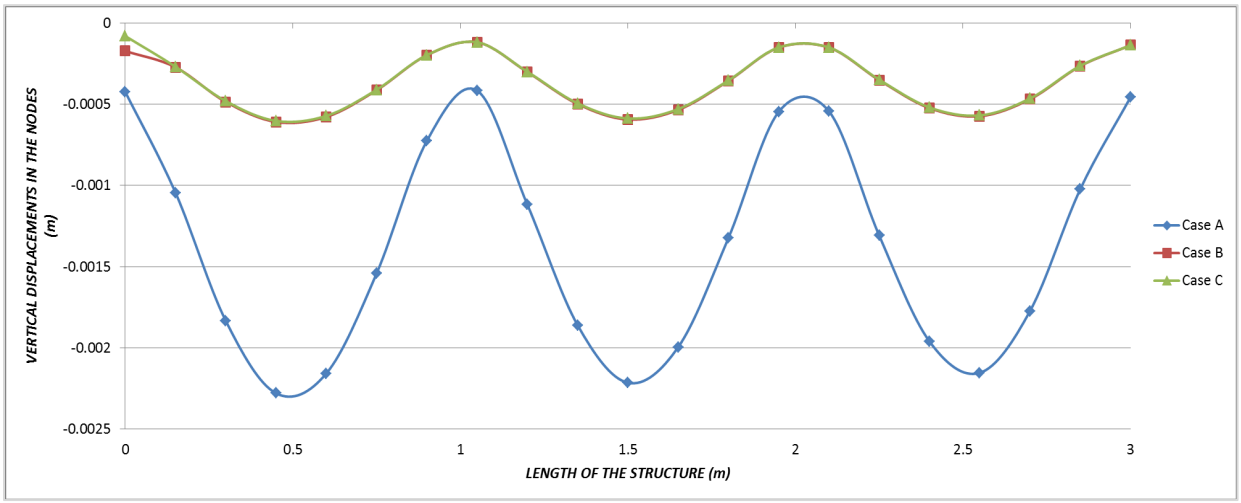


Figure 15. The envelope of vertical displacements in structure type 2.

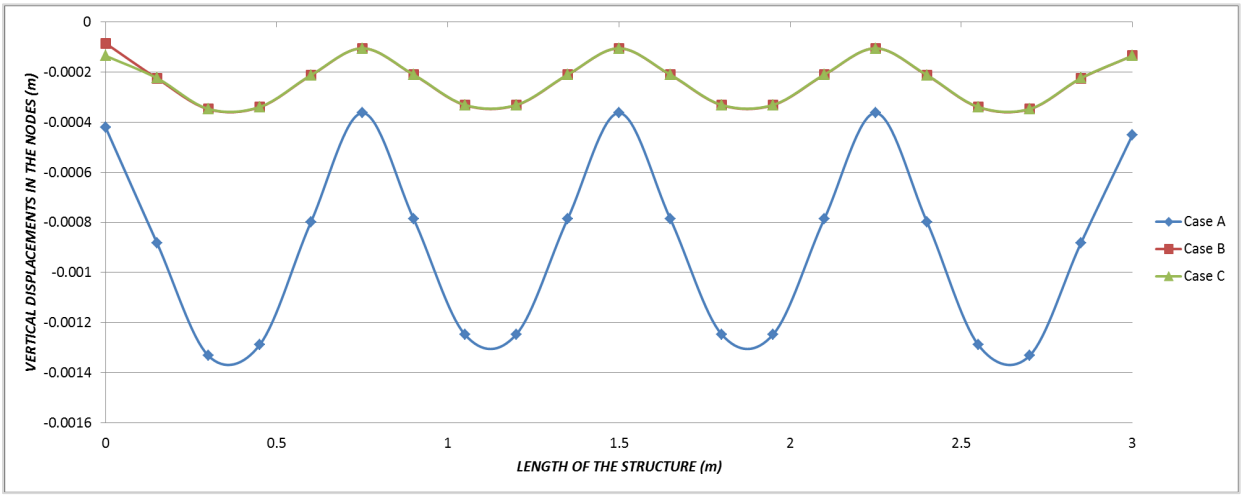


Figure 16. The envelope of vertical displacements in structure type 3.

tendency of its behavior is the same. Negligible deformations in the supports and the highest vertical displacements in the middle of the empty space. Clearly, the quantity of this results shows that vertical displacements descend in the middle until near to 0.001 meters. The deformation is depreciable if is analyzed by traffic loads type A. Structure type 3 could be utile for traffic heavier and could be used in national roads as possible. However, it is necessary for other possibilities, to modify the magnitude of loads and to make the simulation.

It is possible to say that the structure type 3 has the highest efficiency of strengths distribution. The results, showed in the envelope, permit to identify vertical displacements in the critical points in structure type 3 are more than 70% inferiors than the critical points in the structure type 1. The comparison shows the results that were expected because of the number of support the more resistance of all the structure. Clearly, it is important to define the material is plastic and the behavior analyzed was defined as a continuous and homogeneous material.

Structure type 1 shows a behavior with highest deformations in the borders (case A) when the load is applied in the middle of the empty space. Likewise, structure type 2 shows a similar behavior, but in comparison with the type 1 structure its vertical displacements are reduced by around 28%. Structure type 3 demonstrates the best behavior with vertical displacements; it shows a reduction of 41% compared with structure type 2 and 72% compared with structure type 1 in the same conditions of loads.

Furthermore, it is possible to say that is unnecessary to use more supports in the structure, considering the traffic loads type A. It is important to clarify that the material should be homogeneous and the construction should guarantee high quality in the structure in order to get the best behavior in a prototype. Tested in laboratory or in camp tests.

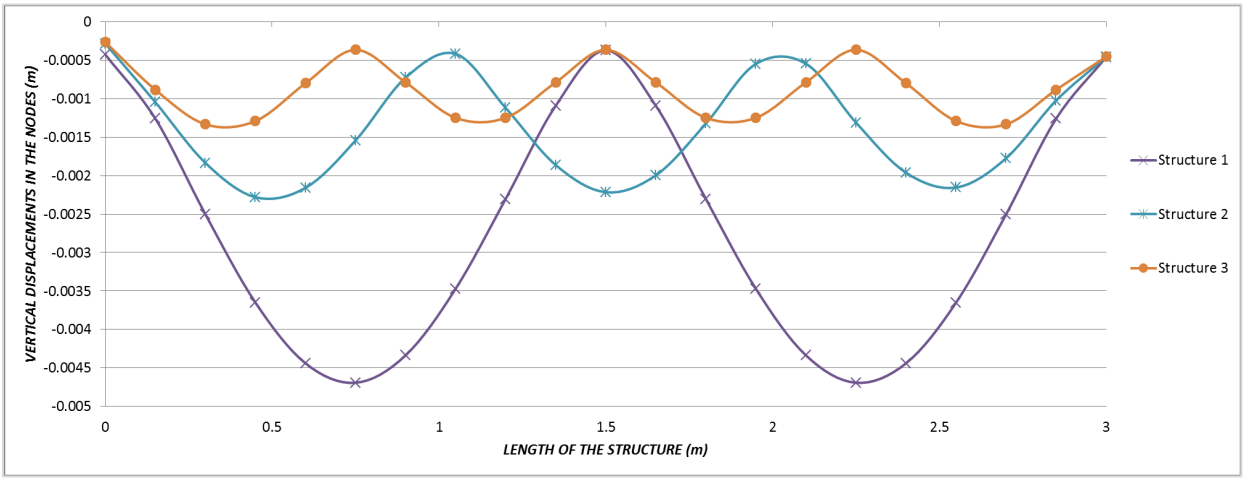


Figure 17. Comparison of envelopes of vertical displacements in structure type 1, 2 and 3.

8. Model of optimization

During the process of simulations, it was possible to identify that the structure type 2 displays the necessary resistance for supporting these loads with vertical displacements in the critical points that is a considerably low value. After the analysis of this material in the laboratory, it is possible to infer that Polyethylene Terephthalate can be deformed 0.03 meters and does not lose its plastic characteristics.

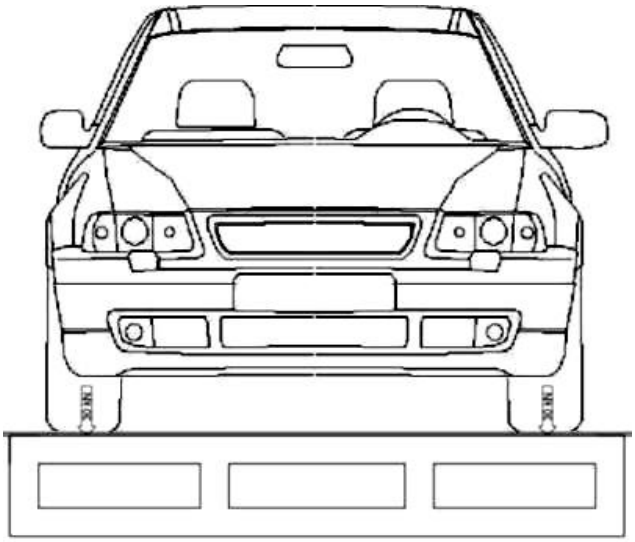


Figure 18. Model of charge for traffic type A.

Furthermore, the comparison of the supports of all vertical displacements show that both types of structures have similar behavior. Consequently, the type 2 structure was selected for developing another simulation, changing the thickness of the structure and reviewing its behavior. Models of simulation proposed one model of charge with one automobile, which induces punctual vertical loads over the structure as is shown in Figure 18. Each load is defined as 30000 N and separated 2.25 meters, according to the description of cars dimensions in the INVIAS manual.

The analysis was developed in the border of the structure (case A). The point where the critical conditions for loads were found. The new simulation was defined with six widths of structure: 0.06, 0.08, 0.10, 0.12, 0.14 and 0.16 meters respectively. Figure 19 shows vertical displacements in the critical points of the structure for the six versions of the model. As was identified in the simulation of critical conditions, it is possible to determine that the structure shows the highest deformation under the point where the loads are applied. This model applied two loads simulating the fires of a car. The central space of the structure showed positive displacements; in that segment the structure permits flector moments to transfer strengths, and the displacements change to the other size of reference ax.

A structure with 0.06 meters of width shows the highest vertical displacement 0.01303 meters. Even though this deformation is high in the simulation, if it is analyzed in real conditions, deformations from 0.013 meters are not very significant; it is possible that one structure has irreversible deformations for the frequency of the load's application.

The structure with 0.08 meters shows deformations of 0.0058 meters. For the 0.10 meters of width, it showed 0.00291 meters. With the structure, 0.12 meters is found 0.00117 meters of vertical displacement. The structure with 0.14 meters of width 0.001001 meters. And finally, for 0.16 meters in simulation, the structure showed 0.000759 meters of deformation in the borders.

Figure 20 contains critical deformations in all the type 2 structures with different widths. The loads were analyzed parallel to traffic direction. It is possible to infer that for applications on a real scale, it is necessary to identify one mechanism for transferring strains.

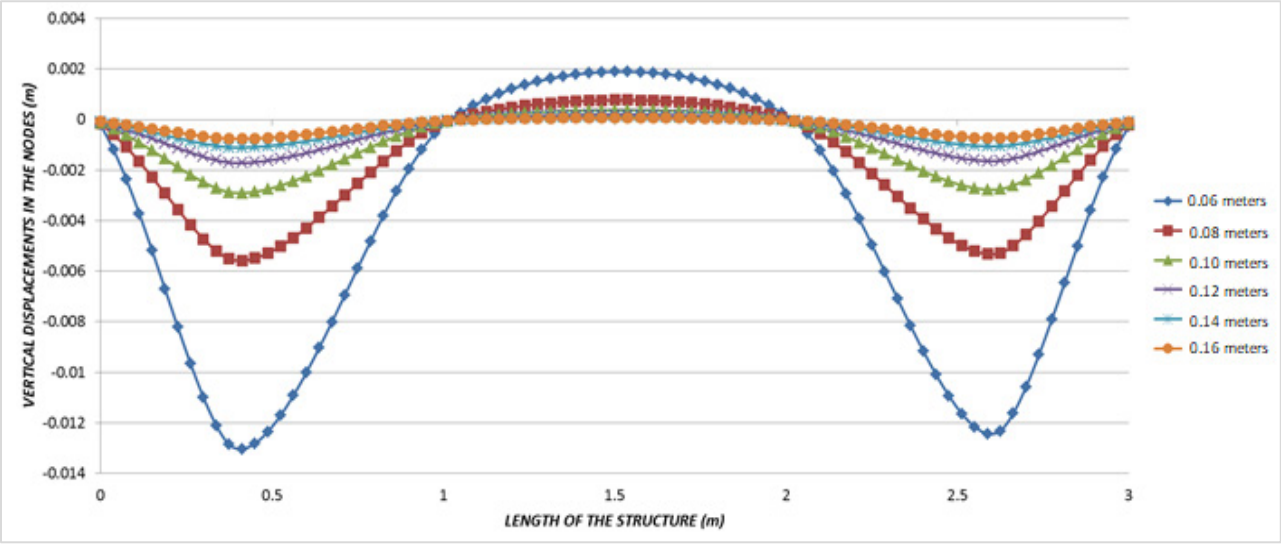


Figure 19. Enveloped of vertical displacements for structure type 2 with different width cross-direction to traffic.

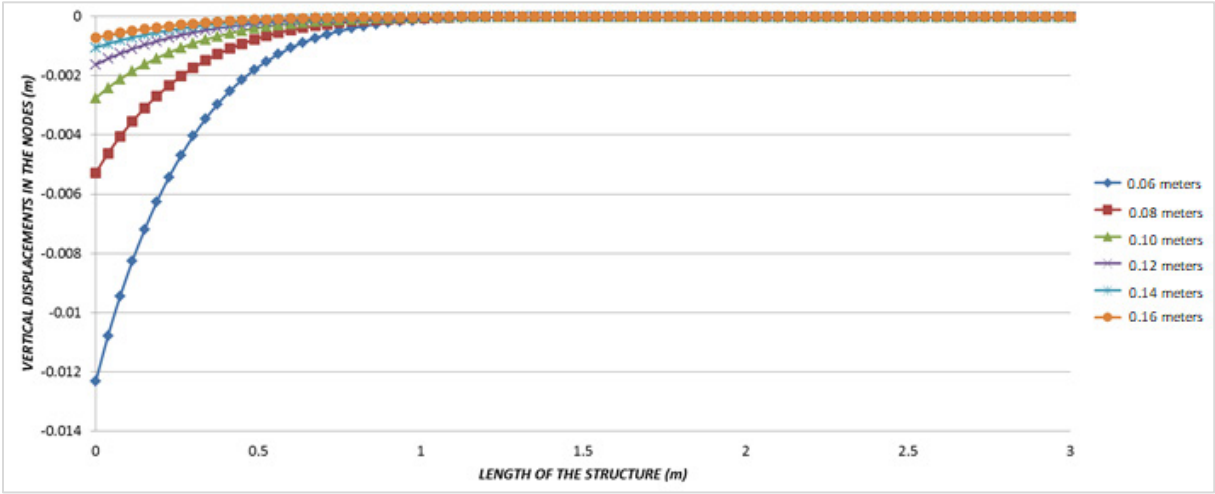


Figure 20. Enveloped of vertical displacements for structure type 2 with different width parallel to traffic direction.

Conclusion

Polyethylene Terephthalate highly pollutes the environment, and is frequently disposed without control in landfills or incinerated. Because of its characteristics this material can be recycled and transformed for a variety of uses while reducing its polluting potential. Polyethylene Terephthalate`s high resistance and flexibility inspire studies into its potential uses. The research proposes developing rolling surfaces in pavement structures.

The simulation offers considerable results according to the identification of different types of structures and different width in one of those structures. Structure type 2 showed important results because the vertical displacements are permissible and it permits optimization of the quantity of material required for making each element. Different widths show the structure supports without difficulties of traffic loads type A and would support loads type B in some cases.

Admissible deformations were the basis for changing the model of structures and the width in the last exercise where it was possible to allow for making test prototypes. The critical case shows potential use in cycle paths, walk paths and small streets for automobile vehicles. However, it is necessary to do more studies in resistance, prototypes and especially in protection against fire.

Empty spaces inside the structures are functional. It is possible to use these spaces for electrical or hydraulic ducts or for another kind of service in the structure.

Case A was the most critical situation in all simulations during the research. These results obey the discontinuity of the element in the borders. The results open new ways for research in how transfer charges from one element to the next. The process, even though it would be for construction, helps research to identify different possibilities for making the structure as a prototype or on a real scale.

Another research possibility is the characteristics of this behavior when the support of all the structure is not rigid. Simulation was developed using soil deformations under the structure. For traffic type A, these deformations are non-considerable, but in a real process of analysis, it is necessary to introduce the most variables as possible.

References.

[1] A. Gutiérrez, "¿Qué es la movilidad?. elementos para (re) construir las definiciones básicas del campo del transporte", Bitácora Urbano Territorial, Universidad Nacional de Colombia, 21(2), 61-74, 2012.

[2] E. Arencibia, J. Junco, M. Perera, "Análisis de la contaminación atmosférica de la planta de asfalto caliente "Abel Santamaria", de coliseo", Revista Avanzada Científica, 16(3), 2013.

[3] F. Reyes, M. Madrid, S. Salas, "Mezclas asfálticas modificadas con un elastómero (caucho) y un plastómero (tiras de bolsas de leche con asfalto 80-100)", Infraestructura Vial, Pontifica Universidad Javeriana, 1(17), 25-34, 2007.

[4] N. Kalantar, R. Karim, A. Mahrez, "A review of using waste and virgin polymer in pavement", Construction and Building Materials, 33, 55-62, 2012.

[5] R. Siddique, J. Khatib, I. Kaur, "Used of recycled plastic in concrete: A review", Waste Management, 28(10), 1835-1852, 2008.

[6] D. Martínez, L. Cote, "Diseño y fabricación de ladrillo reutilizando materiales a base de PET", INGE CUC, 10(2), 76-80, 2014.

[7] J. Carswell, F. Cruz, "Misión y ventajas de los betunes modificados con polimeros", Revista de la Asociación Técnica de carreteras, 1(63), 33-39, 1997.

[8] A. Tellez, "La complejidad de la problemática ambiental de los residuos plásticos: una aproximación al análisis narrativo de política pública en Bogotá", MSc Thesis, Universidad Nacional de Colombia, 2012.

[9] G. Gourmelon, "Global plastic production rises, recycling lags", Vital Signs Online, World Watch Institute, Washington, DC, United States, 2015.

[10] R. Sarria, G. J, "La gran problemática ambiental de los residuos plásticos: microplasticos", Journal de Ciencia e Ingeniería, 8(1), 21-27, 2016.

[11] J. Jambeck, R. Geyer, C. Wilcos, T. Siegler, M. Perryman, A. Andrady, K. Lavender, R. Natayan, K. Lavender, "Plastic waste inputs from land into the ocean", Science, 347(6223), 768-771, 2015.

[12] M. Friksen, L. Lebreton, H. Carson, M. Thiel, C. Moore, L. Borrero, J. Reisser, "Plastic polution in the world's oceans: more than 5 trillon plastic pieces weighing over 250,000 tons afloat at sea", PLoS ONE, 9(12), 2014.
[13] Departamento Nacional de Planeación, "Rellenos sanitarios de 321 municipios colapsaran en cinco años, advierte el DNP", Bogotá, Colombia, 2016.
[14] A. Melro, P. Camanho, F. Andrade, S. Pinho, "Micromechanical analysis of polymer composites reinforced by unidirectional fibres: Part I-Constitutive modelling", International Journal of Solids and Structures, 50(11), 1897-1905, 2013.
[15] A. Black, Urban mass transportation planning, McGraw-Hill, New York, United States, 1995.
[16] S. Al-Salem, S. Lettieri, J. Baeyens, "Recycling and recovery routes of plastic solid waste (PSW): a review", Waste Management, 29(10), 2625-2643, 2009.
[17] C. Maharaj, R. Maharaj, J. Maynard, "The effect of Polyethylene Terephthalate particle size and concentration on the properties of asphalt and bitumen as an additive", Progress in Rubber, Plastics and Recycling Technology, 31(1), 1-23, 2015.
[18] C. Swaptik, A. Tashkant, O. Suganya, "Polyethylene Terephthalate (PET) waste as building solution", International Journal of Chemical, Environment and Biological Sciences, 1(2), 308-312, 2013.
[19] H. Kosmatka, B. Kerkhoff, W. Panarese, J. Tanesi, " Design and Control of Concrete Mixtures", Portland Cement Association, Illinois, United States, 2004.
[20] Instituto Nacional de Vías, "Manual de diseño de pavimentos asfalticos para vias con bajos volúmenes de tránsito", Bogotá, Colombia, 2007.
[21] L. Avila, G. Martinez, C. Barrera, F. Ureña, A. Loza, "Effects on mechanical properties of recycled PET in cement-based composites", International Journal of Polymer Science, 1-6, 2013.
[22] Singiresu, Rao, “The finite element method in engineering”, Butterworth-Heinemann, Chennai, India, 2017.
[23] E. Wilson, "Three-Dimensional static and dinamic analysis of structures", Computer and Structure Inc, Berkeley, United States, 2011.
[24] S. Ahmad, B. Irons, O. Zienkiewicz, "Analysis of thick and thin shell structures by curved finite elements", International Journal for Numerical Methods in Engineering, 2(3), 419-451, 1970.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 3, No. 6, 479-488 (2018)
www.astesj.com
Special Issue on Multidisciplinary Sciences and Engineering

ASTES Journal
ISSN: 2415-6698

Multi-Objective Path Optimization of a Satellite for Multiple Active Space Debris Removal Based on a Method for the Travelling Serviceman Problem

Masahiro Kanazaki[*,1], Yusuke Yamada[1], Masaki Nakamiya[2]

[1]*Division of Aerospace Engineering, Graduate School of System Design, Tokyo Metropolitan University, 190-0015, Japan*

[2]*Department of Aerospace Engineering, Teikyo University, 320-8551, Japan*

ARTICLEINFO

Article history:
Received: 14 August, 2018
Accepted: 12 December, 2018
Online: 26 December, 2018

Keywords:
Multiple Space Debris Removal
Trajectory Optimization
Travelling Serviceman Problem
Evolutionary Algorithm
Multi-objective Optimization

ABSTRACT

Space debris removal is currently a critical issue for space development. It has been reported that five pieces of debris should be removed each year to avoid further increase in the amount of debris in orbit. One approach for the removal of multiple pieces of debris is to launch multiple satellites that can each remove one target debris from orbit. The benefit of this approach is that the target debris can be removed without orbit transition, and thus, the satellite can be developed considering simple satellite mechanics. However, to realize this concept, multiple satellites need to be launched. Another approach is to use one satellite to remove multiple pieces of space debris. This approach can reduce the launch costs and achieve efficient removal of space debris. However, the satellite must change its orbit after the removal of each debris piece, and a technique for optimizing the orbit transition is required. In this study, the latter strategy and developed a satellite trajectory optimization method for efficient space debris removal were focused on. The similarity between the problem of multiple space debris removal and the travelling serviceman problem (TSP) were considered, and the TSP solution involving an evolutionary algorithm (EA) was applied. To improve the efficiency of multiple debris removal, the total radar cross-section (RCS), which indicates the amount of space debris, and the total thrust of the satellite was minimized. The TSP solution method was extended to multiple objectives by coupling it with a satellite trajectory simulation. To evaluate the developed method, a set of 100 pieces of space debris was selected from a database. The results indicated a trade-off between the total RCS and total thrust.

1 Introduction

Parts of rockets, spent satellites, and tools left behind during space operations by astronauts continue to remain in orbit as space debris (”debris”) [1, 2, 3, 4]. Such debris pose a threat to satellites in operation, including the International Space Station as the debris components may collide with these satellites and cause critical damage. Furthermore, when the amount of debris exceeds a certain value, the involved parts may collide with each other, and further increase the amount of debris. This phenomenon is referred to as the ”Kessler syndrome”[5], and a theoretical model has been developed in this respect. The self-crushing phenomenon called ”break up” [6] due to the explosion of residual fuel also increases the amount of debris. Thus, increasing debris is an urgent problem that needs to be solved in future space development, and active debris removal techniques are being studied[7, 8, 9]. To this end, several methods using removal satellites have been studied. For example, the use of extending conductive tethers that connect the removal satellite and a piece of debris, and then eliminate the debris by passing electric current in the Earth magnetic field and providing thrust to the debris by using the generated

[*]Corresponding Author: Masahiro Kanazaki, 6-6 Hino, Tokyo, Japan, kana@tmu.ac.jp

https://dx.doi.org/10.25046/aj030656

Lorentz force [7], has been widely investigated. Another possible approach is to attach thrusters to the debris pieces [9] to change their trajectory. Such methods require the satellite to be powered to rendezvous with the debris; therefore, efficient mission scenarios are required.

From the viewpoint of cost-effectiveness, it is desirable to eliminate multiple debris with a removal satellite. In addition to the technology of the removal satellites, the optimization of the order in which debris is collected and/or captured and disposed should be considered. Although the optimization of the meeting order can be achieved by considering general combination optimization problems such as the travelling salesman problem (TSP), the debris moves non-cooperatively towards the dumping satellite in this multiple debris removal problem. Therefore, it is also necessary to investigate the applicability of the TSP solution. In an actual dumping mission, there are multiple objectives such as maximizing the total size of debris to be discarded while minimizing the total amount of energy required for the removal satellite to change its trajectory. In particular, it is predicted that the sum of the sizes of debris represented by the radar cross-section (RCS) and the sum of the trajectory-changing energies associated with each debris are contradictory.

Therefore, in this research, a general methodology for the TSP expanded as a multi-objective problem was applied to the path optimization problem of satellites for active debris removal to eliminate a plurality of debris. The data of 100 observed debris was taken as an example; solve the multi-objective problem by changing the number of target debris to 2, 3, 4, and 5; and determine the trend of the optimum path.

2 Related Study

2.1 Researches on Active Space Debris Removal

In 2017, there were 15,000 pieces of space debris with a size of 10 cm or more. The number of debris pieces is expected to exceed 28,000 in 2116 and more than 65,000 in 2210. However, assuming that five pieces of debris can be removed annually from 2020, the number of debris in 2116 can be reduced to approximately 20,000; if 20 pieces can be removed, this number can be reduced to approximately 16,000, according to the prediction reported in [10]. Although this suggests that multiple debris removal techniques are required, it is expensive to launch and operate multiple removal satellites. Thus, an active debris removal technique using only one satellite operating with path optimization is required.

Conceptual research on multiple debris removal considering cost reduction was carried out at the NASA Goddard Space Flight Center [11]. Conductive tether satellites were adopted in this exercise, and the effectiveness of removal of the assumed 50 pieces of debris was discussed. The optimum trajectory was calculated, and the weight of the removed debris was evaluated by an estimation formula. The examination was conducted using the TSP solution method. In this study, several case studies focusing on the number of removed debris were examined; however, the required thrust for the removal satellite was not considered in the optimization process and only a single objective was considered.

2.2 Solutions for TSP and TSP-like Problems

The TSP [12], which is a combinatorial optimization problem, is known as a non-deterministic polynomial (NP) problem. To solve the TSP, heuristic methods have been shown to be effective. The problem involves obtaining the shortest travelling distance for a salesman, when the salesman is expected to visit all distributed cities, as shown in Fig. 1. Here, the distance of the path between cities i is d_i. The TSP is one of the most popular problems and an example of evolutionary calculation.

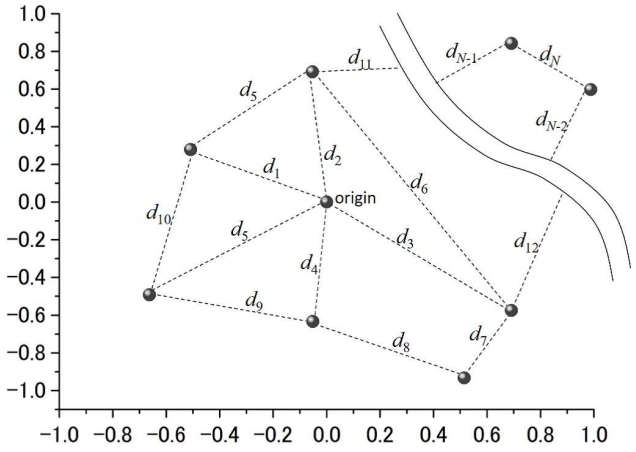


Figure 1: Example of travelling points and routes for the TSP.

A problem similar to the TSP is the "Kyoto Tourism Problem." [13]. When visiting and sightseeing in the city of Kyoto, satisfaction and travel expenses are the factors considered by travelers. The TSP solution method that introduces the concept of the Pareto optimum for the multi-objective problem was applied. In this method, a reciprocal relationship exists between the objectives, and a plan must be proposed based on the travel budget of each traveler.

Another application of the TSP is to reduce a driver's burden and shorten the route for pick up and transfer in consideration of customers and road conditions, which change dynamically [14]. While simulating road conditions and the movement of the customers simultaneously, researchers solved the problem using the evolutionary method. It can be said that this problem is similar to the path optimization problem for active debris removal satellites. However, in reality, it should be noted that debris are uncontrollable

and non-cooperative objects, while customers exhibit cooperative behaviour through communication and self-judgement.

3 Trajectory Optimization for a Multiple Space Debris Removal Satellite

In this research, a plurality of debris is effectively removed by "maximizing the sum total of the RCS of the debris to be removed(RCS_{tot}) and minimizing the sum of the acceleration to move to the ith debris" (ΔV_{tot} which is the amount of ΔV_i for transition).

The debris travels around different orbits (including altitude); thus, it is necessary to increase the speed to meet the orbit in orbit. It is also necessary to minimize the increase in speed to decrease fuel consumption. This problem can be written as follows.

$$\begin{cases} \text{Maximize} & RCS_{\mathrm{tot}} \\ \text{Minimize} & \Delta V_{\mathrm{tot}} \end{cases} \tag{1}$$

Assuming that N_{debri} pieces of the debris are removed, RCS_{tot} can be expressed as follows:

$$RCS_{\mathrm{debri}} = \sum_{i=1}^{N_{\mathrm{tot}}} RCS_i. \tag{2}$$

ΔV_{tot} can be written as the summation of the velocity increment ΔV_i for each debris i:

$$\Delta V_{\mathrm{tot}} = \sum_{i=1}^{N_{\mathrm{tot}}} \Delta V_i. \tag{3}$$

ΔV_i is obtained by solving Lambert's problem [15], which is described in the following subsection.

3.1 Lambert's Problem for Trajectory Evaluation

In this research, the Lambert problem [15] in orbital dynamics was solved to evaluate the path of the removal satellite. This can be rephrased as a problem in which the spacecraft achieves the necessary speed in the orbit to travel from a certain point P to another point Q. The Lambert equation can be expressed as

$$\Delta t = \sqrt{\frac{a^3}{\mu}}[2k\pi + (E - e\sin E) - E_0 - e\sin E_0]. \tag{4}$$

where Δt is the duration of time, a is the semi-major axis, μ is a gravitational parameter, k is the number of revolutions, E is an eccentric anomaly, and e is the orbital eccentricity. ΔV_i is obtained from the target debris position after Δt, the initial position of the debris removal satellite, and Δt is obtained by the expression 4. Figure 2 shows an example of the solution of the Lambert problem translating the trajectory of the removal satellite with respect to the debris.

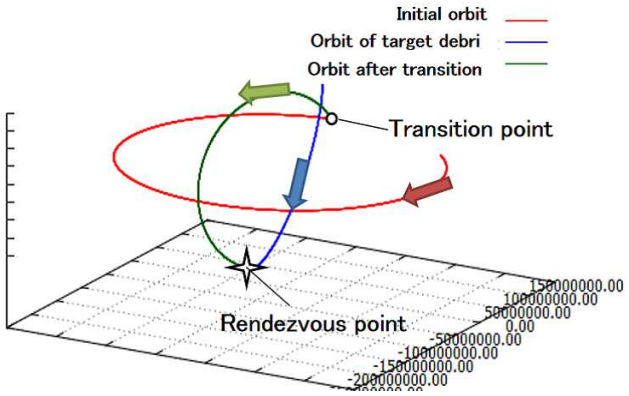


Figure 2: Example of trajectory calculation when the removal satellite heads to the rendezvous point with the debris with one orbit change. Red denotes the original trajectory of the dumped satellite, green denotes the trajectory after the transition, and blue denotes the trajectory of the debris.

4 Combinatorial Optimization and Expansion to Multi-Objective Problem

In this research, the TSP solution methods using a genetic algorithm was applied to the path optimization problem of space debris removal satellites. The following typical algorithm was firstly developed to solve TSP.

- As shown in Fig. 1, it was assumed that all the cities are represented by coordinate points on the plane and N intercity routes are defined.
- A salesman does not visit the same city more than once.

The total travelling distance d_{tot} in Fig. 1 can be written as

$$d_{\mathrm{tot}} = \sum_{i=1}^{N} d_i. \tag{5}$$

4.1 Genetic Operators

4.1.1 Representation of the Combination of the Traveling Path

Path representation (PR) is applied. For example, if the number of visited points is nine, then the order in which the points are visited can be represented as follows::

$$|\,a\,|\,c\,|\,b\,|\,g\,|\,h\,|\,i\,|\,d\,|\,f\,|\,e\,|. \tag{6}$$

4.1.2 Selection

A roulette selection is applied to select individuals for the crossover and mutation operations. In the roulette selection, the selection probability p_i can be based on fitness f_i from N_{pop}:

$$p_i = \frac{f_i}{\sum_{k=1}^{N_{\text{pop}}} f_k}. \tag{7}$$

4.1.3 Crossover

Order crossover (OX) [12] is applied to two selected individuals. In OX, the individuals p_1 and p_2 are selected first:

$$\begin{aligned} p_1 &= |\,a\,|\,b\,|\,c\,|\,d\,|\,e\,|\,f\,|\,g\,|\,h\,|\,i\,| \\ p_2 &= |\,d\,|\,a\,|\,b\,|\,h\,|\,g\,|\,f\,|\,i\,|\,c\,|\,e\,|. \end{aligned} \tag{8}$$

Then, the parts of the gene sequence that are copied to children c_1 and c_2 are selected from p_1 and p_2, respectively. An example is shown in Eq. 9. Here, fourth–seventh genes are copied in that order to the children, and the rest of the genes —which are temporarily represented by $|\,*\,|$— are undecided gene sequence parts to be determined by a crossover:

$$\begin{aligned} c_1 &= |\,*\,|\,*\,|\,*\,|\,d\,|\,e\,|\,f\,|\,g\,|\,*\,|\,*\,| \\ c_2 &= |\,*\,|\,*\,|\,*\,|\,h\,|\,g\,|\,f\,|\,i\,|\,*\,|\,*\,|. \end{aligned} \tag{9}$$

To determine the genes temporarily represented by $|\,*\,|$ of c_1, the following gene sequence of p_2 that corresponds to the undecided gene sequence part of c_1 is copied. The sequential order of this gene sequence remains in the clockwise direction:

$$p_{2\text{b}} = |\,c\,|\,e\,|\,d\,|\,a\,|\,b\,|\,h\,|\,g\,|\,f\,|\,i\,|. \tag{10}$$

Here, the gene sequence part that has already been copied from p_1 to c_1 is removed from $p_{2\text{b}}$:

$$p_{2\text{c}} = |\,e\,|\,d\,|\,b\,|\,h\,|\,i\,|. \tag{11}$$

c_1 can be determined by inserting $p_{2\text{c}}$:

$$c_1 = |\,\underline{b}\,|\,\underline{h}\,|\,\underline{i}\,|\,d\,|\,e\,|\,f\,|\,g\,|\,\underline{e}\,|\,\underline{d}\,|. \tag{12}$$

c_2 can be determined in the same way:

$$c_2 = |\,\underline{a}\,|\,\underline{b}\,|\,\underline{c}\,|\,h\,|\,g\,|\,f\,|\,i\,|\,\underline{d}\,|\,\underline{e}\,|. \tag{13}$$

4.1.4 Mutation

Mutation is used to maintain population diversity, creating individuals with genetic information that cannot be generated only by a crossover. In this research, an inversion mutation that exchanges the genetic information at a position determined by a random number is used. In the inversion method, when p_1 shown in Eq. 8 is chosen as a mutation target, two genes are arbitrarily selected and the positions are exchanged as follows.

$$c_1 = |\,a\,|\,b\,|\,\underline{e}\,|\,d\,|\,\underline{c}\,|\,f\,|\,g\,|\,h\,|\,i\,|. \tag{14}$$

In this example, the third and fifth genes of p_1 are exchanged to create c_1.

4.2 Investigation of the Developed Method with the Typical TSP

As a verification method, a typical TSP (minimizing the total path distance) was solved with 10 randomly distributed cities and the departure point of a salesman. Eight-hundred generations were executed by setting 100 individuals for each generation. It is assumed that the route between the cities is given by a Euclidean distance and the salesman returns to the starting point.

Figure 4 shows the history of the solution (total route distance) obtained by the method outlined in the previous section. From this figure, the total path distance decreases with each generation and converges at approximately 300 generations. An example of the best individual route obtained for each generation is shown in Fig. 1. Figure 4(a), which is the initial generation, has a waste route. Whereas, in Fig. 1(b)-(d), it can be seen that several paths considered to be efficient are being searched. In particular, Fig. 1(c) and (d) show almost the same evaluation value, but the routes are different. This shows that various solutions can be obtained in the design variable space. In this research as well, designers can select the route according to the mission requirements and constraints using the information on the optimum route for the debris removal satellite.

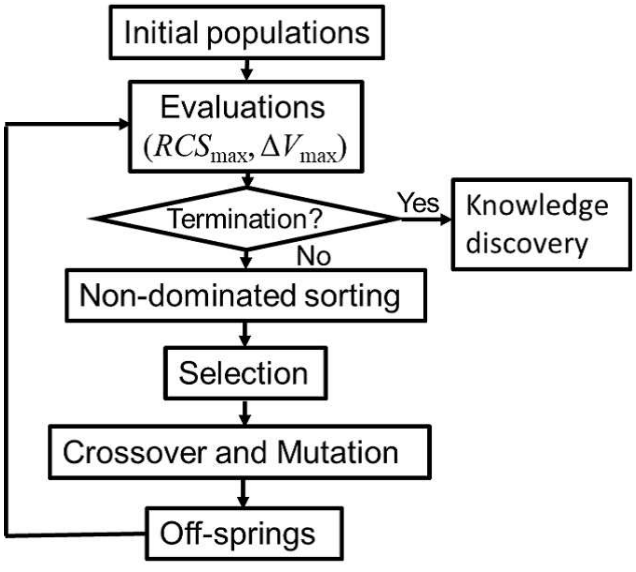


Figure 3: Procedure of multi-objective trajectory optimization of a satellite for multiple active space debris removal.

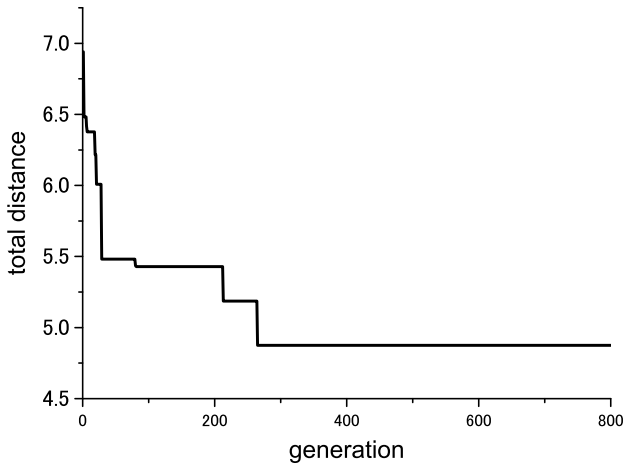


Figure 4: Convergence history of solutions with the developed algorithm

4.3 Non-dominated Sorting for MoPs

The final goal of this research is multi-objective optimization of the orbits of multiple debris removal satellites. For multi-objective optimization, ranking is performed by non-dominated sorting (Fig. 6), which was introduced in the Non-dominated Sorting Genetic Algorithm-II [16]. An elite strategy was adopted to archive good solutions and take over to the next generation without genetic operations such as crossover or mutation.

4.4 Scatter Plot Matrix (SPM)

The solution and design space of the multi-variable design problem obtained by MOEA are observed using SPM [17], which is one of the data mining methods. SPM arranges two-dimensional scatter plots such as a matrix among the objective functions and design variables and facilitates the investigation of the design problem. Each of the rows and columns is assigned attribute values such as design variables, objective functions, and constraint values. The diagonal elements show the same plots. Therefore, it can be said that the SPM shows scatter plots on the upper triangular part of the matrix and the correlation coefficients on the lower triangular part as additional information. modeFrontier™4.2.2 was employed in this study.

5 Results for the Trajectory Optimization of the Multiple Active Space Debris Removal Satellite

In this research, debris data obtained when China destroyed the satellite "Fengyun-1C" in an experiment for the development of an anti-satellite weapon method in 2007 [6] was used. In this experiment, the over 2800 pieces of satellite debris were obtained debris, which is the highest number to date.

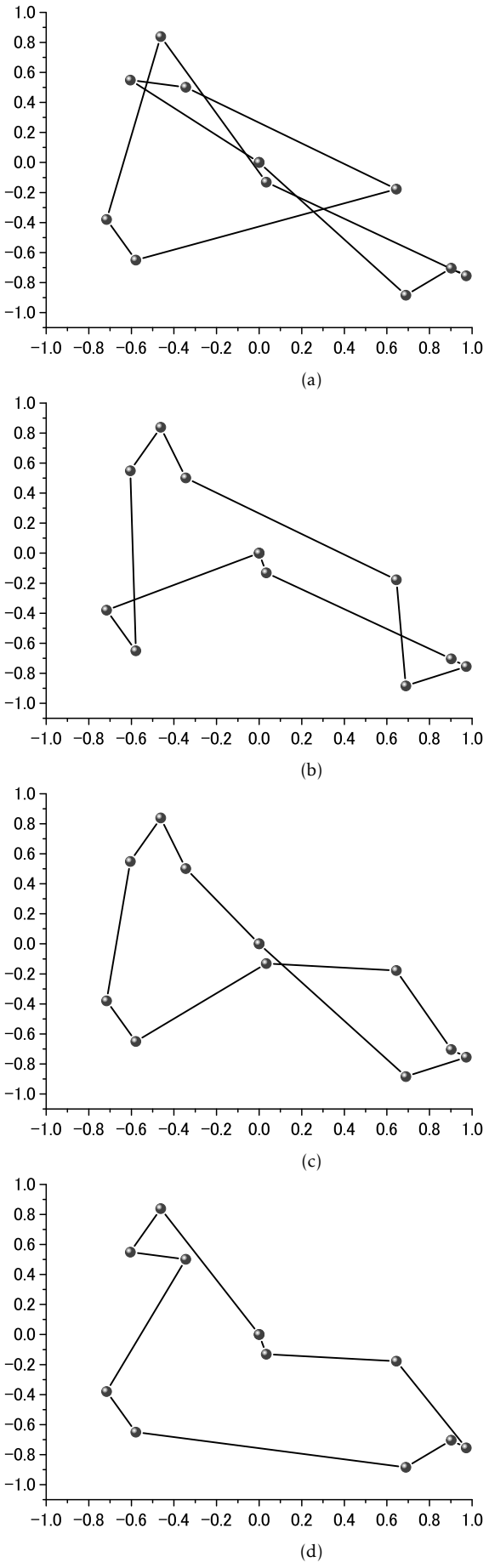


Figure 5: Selected solutions: (a) initial generation, (b) 100th generation, (c)400th generation, and (d) 800th generation

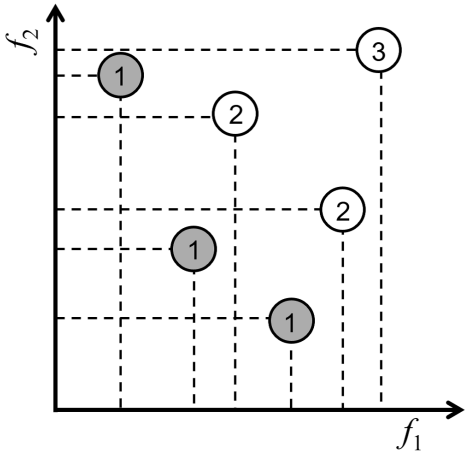


Figure 6: Ranking by the non-dominated sorting algorithm.

The orbital altitude of this debris cloud is located at a high altitude of 800 km; hence, the suspension time is long and is expected to have a negative impact/influence in the long term. Because each piece of debris was tracked from the start, 100 individual debris data which were randomly selected from the catalogue (data table) was used. The frequency distribution of the RCS considered in this research is shown in Fig. 7. RCS_i is based on the observation data published by the North American Aerospace Defence Command.

For optimization, the multi-objective problem shown in the expression 1 in the section 3 was solved. To obtain information on removal efficiency, four cases in which the number of debris to be removed was changed to 2, 3, 4, and 5 were solved. It is assumed that the removal satellite has been in a parking orbit at an altitude of 200 km and departed at 0 o'clock on January 1, 2015. Upon reaching each debris, the satellite conducts a removal operation for 1 h and remains stationary in that orbit until it has to depart for the next debris. Evolutionary calculation was carried out with 100 individuals per generation and 2000 generations were executed.

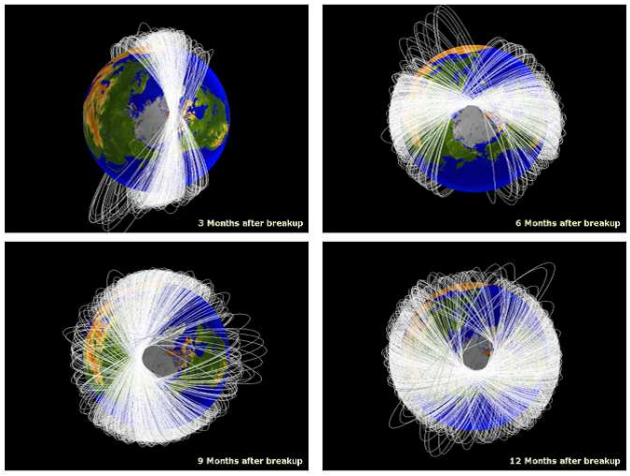


Figure 7: Three months of tracking data for space debris from the satellite destruction test of Fengyun-1C [6].

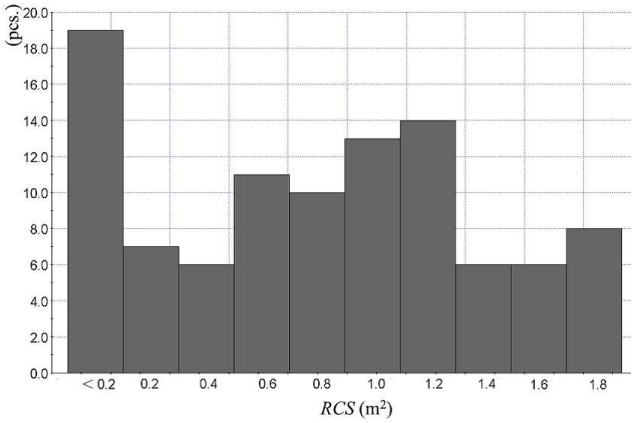


Figure 8: Frequency distribution of the RCS for 100 pieces of debris that were candidates for removal.

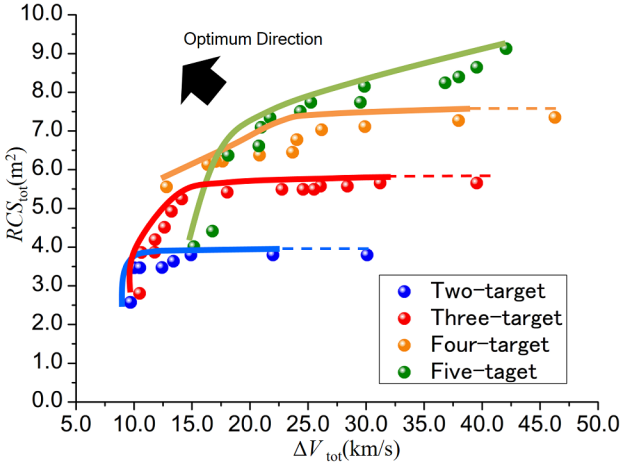


Figure 9: Non-dominated solutions obtained by the developed trajectory optimization method.

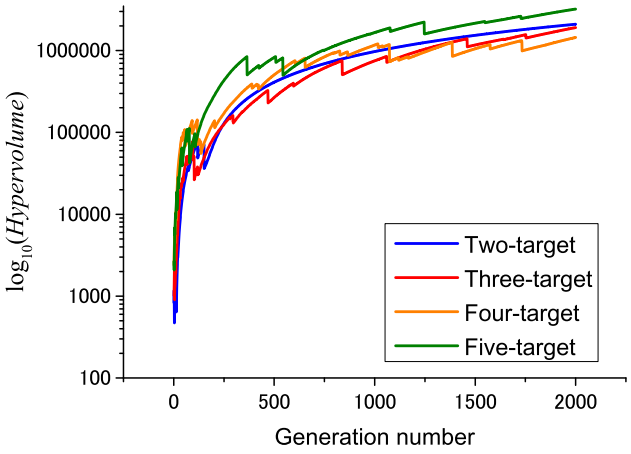


Figure 10: Comparison of hypervolume histories.

5.1 MoP Solutions by Means of MOEA

Figure 9 shows the solution after the 2000th generation. In this figure, the solutions of Rank 1 and 2 were plotted. From Fig. 7, the maximum RCS of one debris is $1.8m^2$; thus, the maximum value of RCS_{tot} in cases of removing two, three, four, and five pieces of debris is $3.6m^2$, $5.4m^2$, $7.2m^2$, and $9.0m^2$, respectively. In each case shown in Fig. 9, the resulting RCS_{tot} shows the maximum possible value. In the figure, the solid line connects Rank 1, and the dotted line connects Rank 2. According to these lines, ΔV_{tot} was distributed over a wide range while RCS_{tot} shows the same maximum values.

The hypervolume convergence history is shown in Fig. 10. In each case, converged solutions were obtained, while it is expected to moderately improve after 2000 generations. This is because the number of solutions constituting the non-dominated solution will increase in each case, while the maximum point of RCS_{tot} has already been determined.

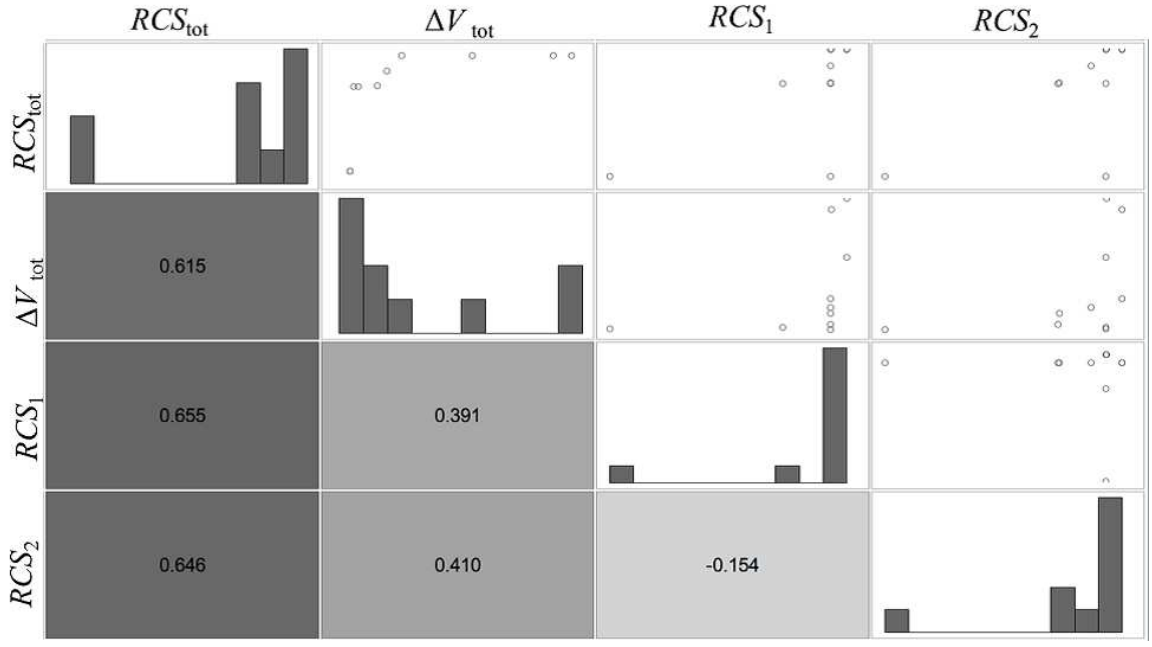
5.2 Visualization of the Design Problem with SPM

Figure 11 shows the SPM for each case. The SPM contains two objective functions and the RCS of the selected debris in the rendezvous order. In Fig. 11(a), both RCS whose debris to be removed show high correlation with both the objective functions. In Fig. 11(b) and (c), the RCS of the last debris to be removed (RCS_2 And RCS_4) is also correlated with both the objective functions. On the other hand, in Fig. 11(d), the RCS (RCS_3) of the third debris to be removed is correlated with each objective function; however, the RCS of the last debris to be removed is not correlated with each objective function.

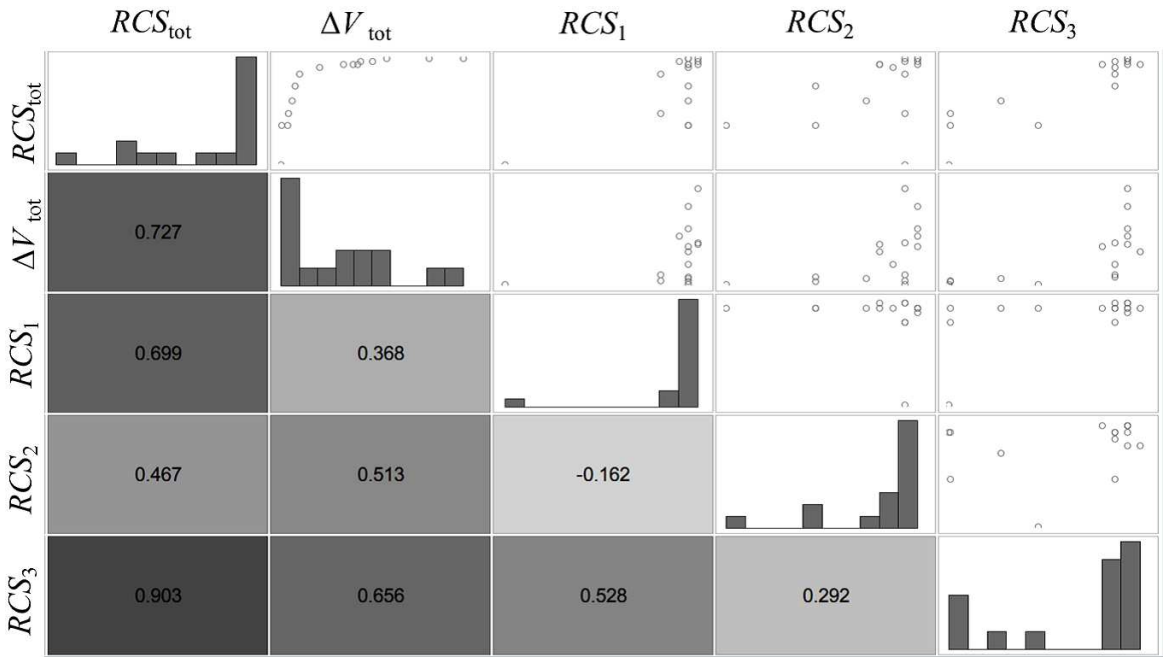
The diagonal elements in Fig. 11 show frequency distributions. Comparing these distributions with those in Fig. 7 shows the deviation. Looking at the frequency distribution of RCS_3 in Fig. 11(d), it was found that relatively large debris (RCS over $1.4m^2$) was removed. Furthermore, as shown in Fig. 11(a) and (b), large debris were selected for RCS_1 and RCS_2, respectively. However, in Fig. 11(c) and (d), it was found that small debris $0.4m^2$ or less were also removed. This means that large debris should be removed along with small debris at a convenient position from the point of orbit transition to improve the overall efficiency. Figure 11(d) shows that in the case of removing five pieces of debris, there is a strong correlation between RCS_{tot} and RCS_3. In fact, when RCS_3 is large, RCS_{tot} is also large. On the other hand, when RCS_3 is large, ΔV_{tot} does not necessarily become large. In this way, it is expected that planning multiple debris removal paths will become easier if the order in which to remove larger debris can be decided.

6 Conclusions

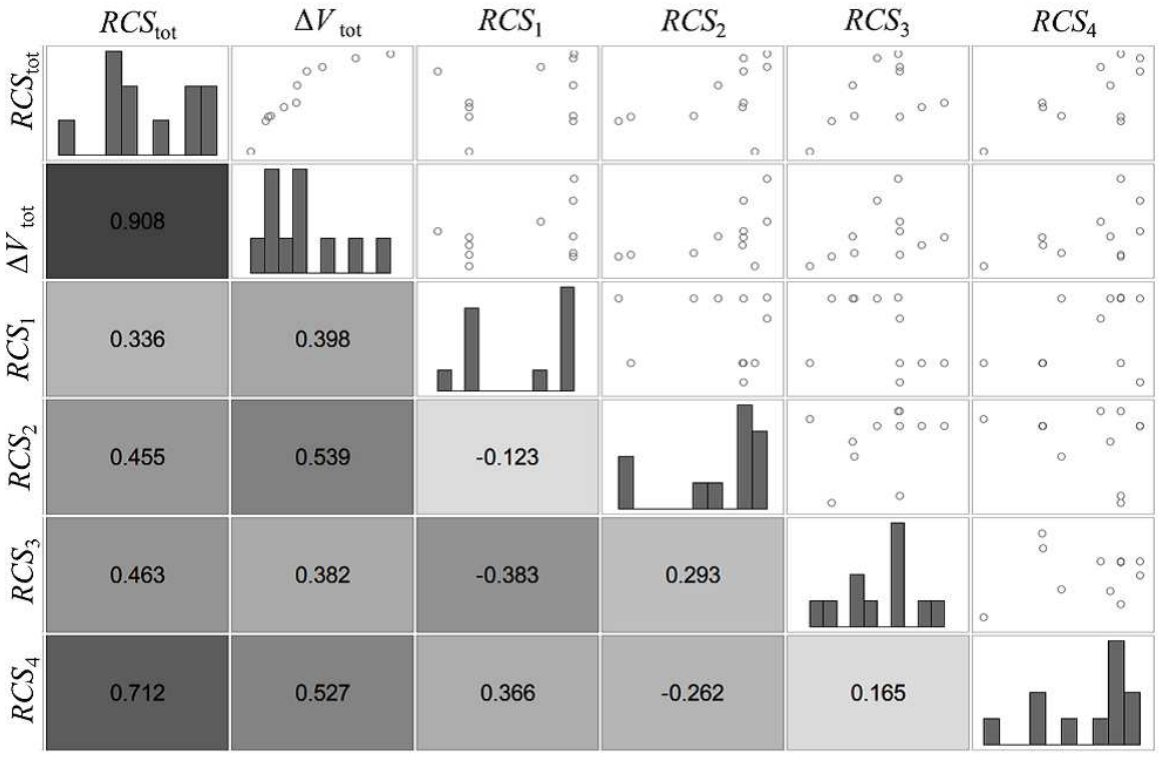
In this research, the path optimization of multiple space debris removal satellites was attempted by applying the TSP solution by using an evolutionary calculation method. The optimization problem considered was the maximization of the sum total of the radar reflection area that represents the size of the debris and minimization of the total velocity increment to rendezvous with the selected debris. Data, including altitude, for hundred pieces of debris were used to investigate our methodology. The Lambert problem was solved for the orbit transition of the removal satellite. In the optimization, order crossover was used as a crossover method. The reverse order method was



(a)



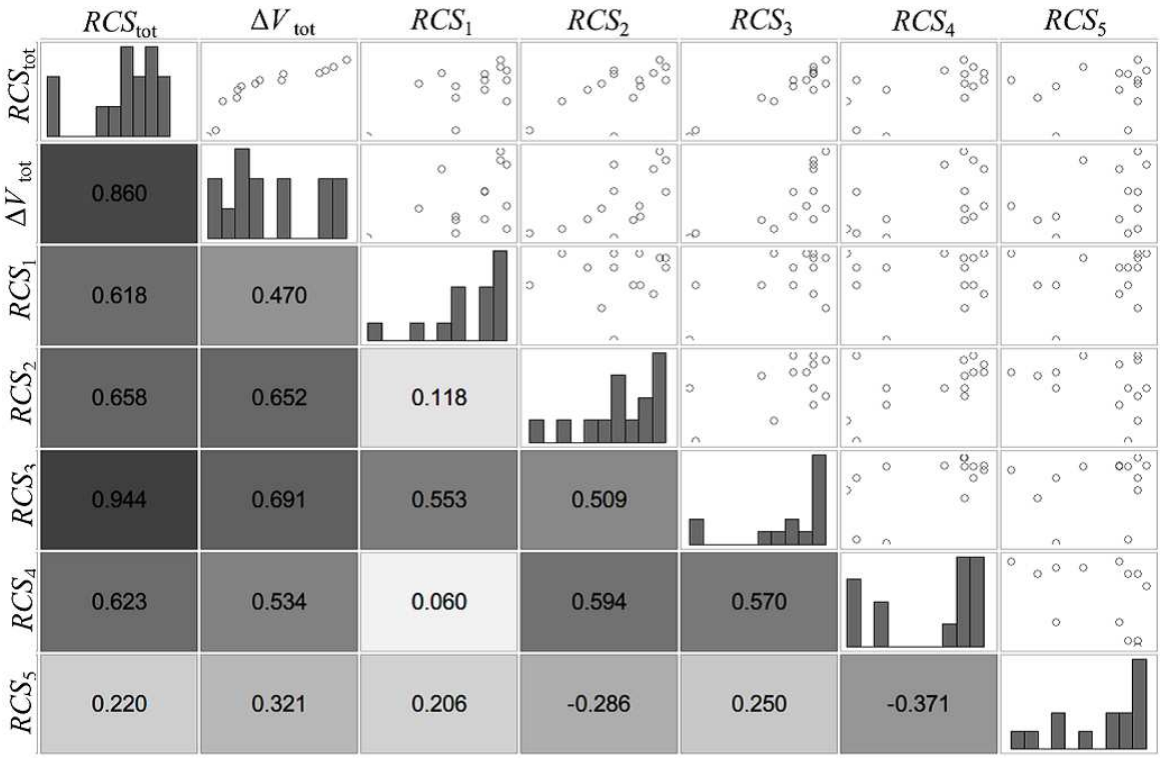
(b)



(c)

used for mutation. Furthermore, as the problem in this study is defined as multi-objective, non-dominated sorting for ranking was employed . In investigating the developed method, multi-objective problems were used by changing the pieces of debris removed by one removal satellite from 2 to 5. According to results, the following information was obtained.

- Because the sum of the maximum radar reflection areas can be calculated from 100 debris candidates in the proposed optimization process, it is effective to use TSP solution method in this problem.
- There is a trade-off between the sum of the sizes of debris to be removed and the total velocity increment of the removal satellite.
- Even when the sum of the sizes of debris to be removed is the maximum, the total velocity increment was not uniquely determined. Thus, the total velocity increment should be set as the objective function.



(d)

Figure 11: Design problem visualization by SPM: removal of (a) two, (b) three, (c) four, and (d) five pieces of debris.

- By increasing the pieces of removed debris simultaneously, the removal operation can be effective because smaller debris will be removed along with larger debris.
- In the case where five pieces of debris were removed, a positive correlation was observed between the radar reflection area of the third piece of debris and the sum of the radar reflection areas. Such findings can be considered as useful knowledge for mission planning.

References

[1] N. N. Smirnov, Space Debris: Hazard Evaluation and Debris, ESI Book Series, Taylor and Francis Group, 2002.

[2] H. Klinkrad, P. Beltrami, S. Hauptmann, C. Martin, H. Sdunnus, H. Stokes, R. Walker, and J. Wilkinson,"The ESA Space Debris Mitigation Handbook 2002," Advances in Space Research, Elsevier, **34**(5), 1251–1259, 2004. https://doi.org/10.1016/j.asr.2003.01.018

[3] T. Yasaka, R. Hanada, and H. Hirayama, "Geo debris environment: A model to forecast the next 100 years," Advances in Space Research, Elsevier, **23**(1),191–199, 1999. https://doi.org/10.1016/S0273-1177(99)00004-6

[4] D. J. Kessler and B. G. CourPalais, "Collision frequency of artificial satellites: The creation of a debris belt," Journal of Geophysical Research, **83**(A6), 2637-2646, 1978. https://doi.org/10.1029/JA083iA06p02637

[5] Kessler, D. J., and Cour-Palais, B. G., : Collision frequency of artificial satellites: The creation of a debris belt, *Advances in Space Research*, Elsevier, pp. 2637–2646 (1978)

[6] Johnson, N. L. Stansbery E., Liou, J. C., Horstman, M., Stokely, C. and Whitlock, D., "The characteristics and consequences of the break-up of the Fengyun-1C spacecraft," Acta Astronautica, Elsevier, **63**, 149–156, 2013. https://doi.org/10.1016/j.actaastro.2007.12.044

[7] V. Aslanov and V. Yudintsev, "Dynamics of large space debris removal using tethered space tug," Acta Astronautica, Elsevier, **91**, 149–156, 2013. https://doi.org/10.1016/j.actaastro.2013.05.020

[8] Bombardelli, C. and Pelaez, J., : Ion Beam Shepherd for Contactless Space Debris Removal, *Advances in Space Research*, Elsevier, Vol, 34, No. 3, pp. 916–920 (2011)

[9] L. T. DeLucaa, F. Bernellia, F. Maggia, P. Tadinia, C. Pardinib, L. Anselmob, M. Grassic, D. Pavarind, A. Francesconid, F. Branzd, S. Chiesae, N. Violae, C. Bonnalf, V. Trushlyakovg, and I. Belokonov, "Active space debris removal by a hybrid propulsion module," Acta Astronautica, Elsevier, **91**, 20–33, 2013. https://doi.org/10.1016/j.actaastro.2013.04.025

[10] Payne, T. and Morris, R., : The Space Surveillance Network (SSN) and Orbital Debris, *33rd Annual AAS Guidance and Control Conference*, AAS Paper Number 10-012, (2010)

[11] Barbee, B. W., Alfano, S., Pinon, E., Gold, K. and Gaylor, D., : Design of Spacecraft Missions to Remove Multiple Orbital Debris Objects, *35rd Annual AAS Guidance and Control Conference*, AAS Paper Number 12-017, (2012)

[12] D. B. Fogel, "An evolutionary approach to the travelling salesman problem," Biological Cybernetics, Springer, **60**(2), 139–144, 1988. https://doi.org/10.1007/BF00202901

[13] K. Kondo, K. Miki, and T. Hiroyasu, "Kyoto touring problem - proposal of a new discrete test problem in multiobjective optimization -," in IPSJ Symposium, Tokyo, Japan, 2003. http://jglobal.jst.go.jp/en/public/200902275803303852 (in japanese)

[14] J. Xiao, X. Ma, Z. Zhu, J. Zhou, and Y. Yang, "Multiobjective memetic algorithm for solving pickup and delivery problem with dynamic customer requests and traffic information," in 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada, 2016. https://doi.org/10.1109/CEC.2016.7744028

[15] D. D. Mueller, R. R. Bate, and J. E. White, Fundamentals of Astrodynamics, New Dover Publications, 1971.

[16] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Transactions on Evolutionary Computation, **6**(2), 182–197, 2002. https://doi.org/10.1109/4235.996017

[17] R. A. Becker, W. S. Cleveland, and A. R. Wilks, "Dynamic graphics for data analysis," Statistical Science, Institute of Mathematical Statistics, **2**(4), 355–383, 1987. https://www.jstor.org/stable/2245523

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 6, 489-495 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Perfect Molding Challenges and The Limitations "A Case Study"

Tan Lay Tatt[*,1], Lim Boon Huat[1], Rosli Muhammad Tarmizi[2], T. Joseph Sahaya Anand[3]

[1]*Infineon Technologies (Malaysia) Sdn Bhd, Engineering Department, 75350, Malaysia*

[2]*Infineon Technologies (Malaysia) Sdn Bhd, Engineering Department, 75350, Malaysia*

[3]*Universiti Teknikal Malaysia Melaka (UTeM), Faculty of Manufacturing Engineering, 76100, Malaysia*

ARTICLE INFO

Article history:
Received: 12 September, 2018
Accepted: 16 December, 2018
Online: 23 December, 2018

Keywords:
Mold Void
Temperature Control
Pressure Control
Transfer Mold

ABSTRACT

Driven by today's market demand, semiconductor is pushing towards the zero-defect direction. The improvement demanded in semiconductor manufacturing is becoming increasingly challenging. In this paper, common molding defects comprise of voids, incomplete fills, and piping holes are studied systematically, focusing on three key areas: 1) Potential mold flow weakness; 2) Molding temperature stability; as well as 3) Defined pressure effects. The in-depth understanding of mold flow in the LF design is achieved via mold flow numerical tool. The numerical model prediction is verified by short shots and end-of-line auto vision data. Advance Process Control (APC) is adopted to measure the stability of key molding parameters like temperature, transfer profile and pressure. The mechanism of transferring the compound in relation to pressure is also analyzed and its effect to molding quality is also assessed. A methodical approach is utilized to understand the process and equipment built-in capabilities from two different equipment manufacturers. The real time transfer profile monitoring is activated for diagnosis of the system issue which leads to the finding of design error of a critical component. The dual temperature controller on one of the systems is analyzed to stabilize temperature for improved compound flow-viscosity control. The process limitations are assessed and transfer profiles are optimized to modify the melt front. By shifting the molding defects to non-critical location, the formation of void at the 500um diameter bonded wire loop peak will be avoided. The verification of potential negative impacts resulted from changes to improve voids, incomplete fills and piping holes are also included in this study. Up-front analysis by adopting numerical tool as a means of understanding the existing design and identifying improvement approach are proven to be useful.

1. Introduction

As today's market requirement on semiconductor is moving towards the direction of zero defect. This demanding requirement has led to improvement not only at manufacturing process control but also at the mold design level without additional cost. The molding process is being chosen due to its high through put per cycle of production. Any imperfection of this process will lead to high quantity of rejections. Ishikawa diagram, Pareto analysis, and Taguchi design of experiments tools are often adopted to address the molding related issues [1,2] but lack of online feedback data to ensure sustainable quality becomes questionable. A computer-aided mold design tool with mold shot library built-in to reduce lead time and cost of mold tool design and fabrication has been formulated [3]. This developed methodology does not take into consideration the interaction effects of mold compound, process parameters, and design. In recent years, vacuum assisted molding has been accepted as a potential solution to reduce void formation [4,5]. However, this feature will incur higher running cost. The surge of introduction of Cu wire to replace Au wire due to high Au price [6] has leaded to new sets of bonding metallurgy and process issues. The effects are pre-mature failures and higher production yield losses [7-11]. Adversely, Cu wire brings along molding advantage of resistant to wire sweep due to its higher modulus. Thicker diameter wire can also provide high resistance to wire sweep during filling of mold cavities [12]. But, it can also

[*]Corresponding Author: Tan Lay Tatt, Infineon Technologies (Malaysia) Sdn Bhd, Contact No: +6062325266 & Email: laytatt.tan@infineon.com

influence the compound flow which can trap the air and cause mold voids.

In this paper, the potential pitfall of molding due to compound flow, temperature stability, pressure control, static DOE results, and interaction effects are being investigated using numerical tool, advance process data control, and in-depth understanding of equipment design.

2. Potential mold flow weakness

The design of package had played the main role that can influence the mold flow pattern [12]. Design of package included die size, wire size, number of wire, wire bonding pattern…etc. Figure 1 showed the DOE matrix for simulation, by using the Autodesk Moldflow.

DOE Matrix	Device #1	Device #2	Device #3
Die Size	Thin	Medium	Thick
Quantity of thin wire *same wire size	1	1	1
Quantity of thick wire *same wire size	3	4	4

Figure 1. Experiment Matrix on different bonding Configuration

Simulation tool with time lapse comparison shows that, the completion of encapsulation process gives faster for bigger die size or higher wire density package as shown in figure 2.

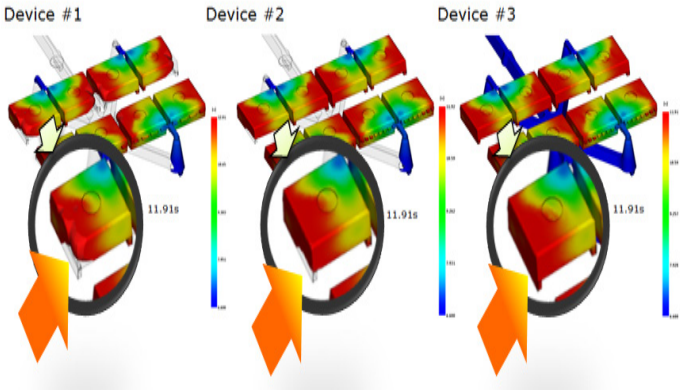


Figure 2. Experiment result on different bonding configuration

This had explained that with constant control on material (compound types), process method and process parameter (cavity temperature, transfer pressure, and transfer profile), package with bigger die size or higher wire density will experience the continuous transfer pressure from plunger after complete fill up the cavity [13]. Higher compactness on package occurs once the compound has been compressed with the set transfer pressure.

3. Temperature Control Sensitivity

Temperature control is one of dominant factors that can greatly affect molding quality [14]. Figure 3 showed the failure mode of mold void detected during production. The magnitude of mold void losses are significantly different by equipment types. "Equipment A" with pin gate concept shows much higher losses as compared to "Equipment F" which is standard mold concept as shown in Figure 4.

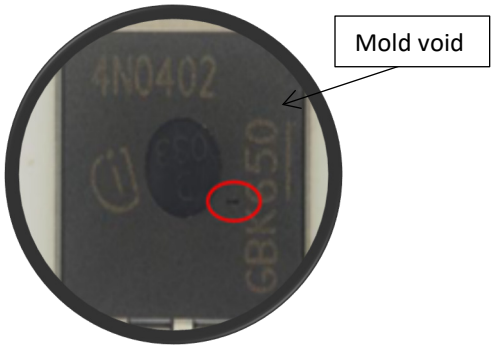


Figure 3. Mold void defect mode

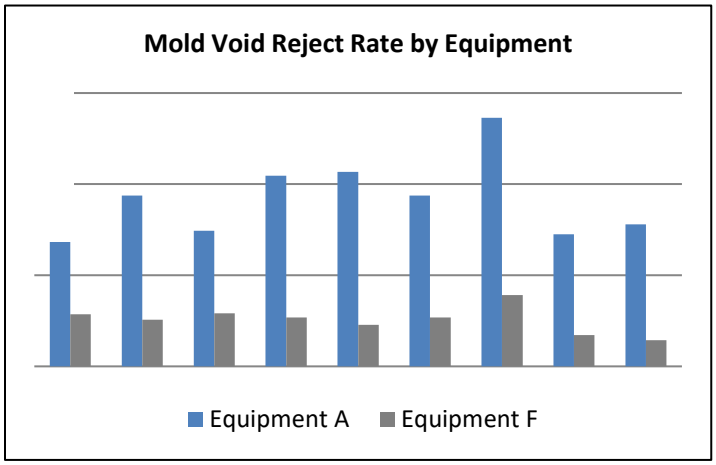


Figure 4. Mold void trend by equipment

Investigation shows that the temperature monitoring trend for "Equipment A" is instable compared to "Equipment F" as shown in Figure 5.

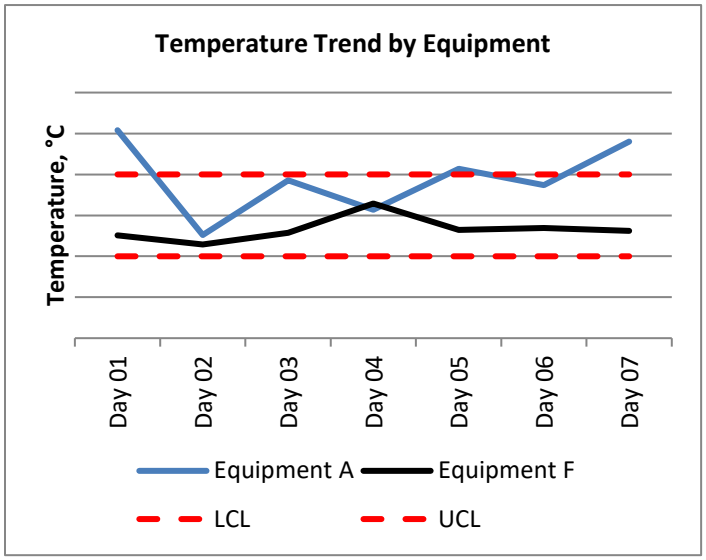


Figure 5. Temperature monitoring Trend

To understand the behaviour of the temperature fluctuations, three areas of focus are defined:

1) Mold design;

2) Heater system; and

3) Robustness of Thermocouple wires.

Mold designs of these two types of equipment are assessed. Figure 6 showed the standard molding concept. The mold runner flow across the rows through the leadframe. The runner which is removed later at the degate station together with the cull. For the standard mold design, mold cavity temperature is stable and sustainable as top cavity and bottom cavity are same temperature setting on solid piece of metal.

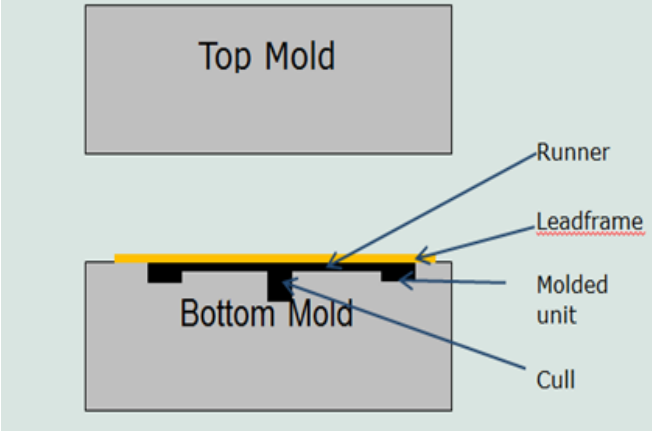


Figure 6. Standard mold design

Figure 7 shows pin gate molding concept. The mold runner flow through a middle plate, middle plate is a moving part whereby it will de-cull the runner and cull during mold open. Due to the middle plate which is moving part and no heater element attached on it, the temperature is always unstable with +/- 5 degree fluctuation caused by heat losses during mold process cycle.

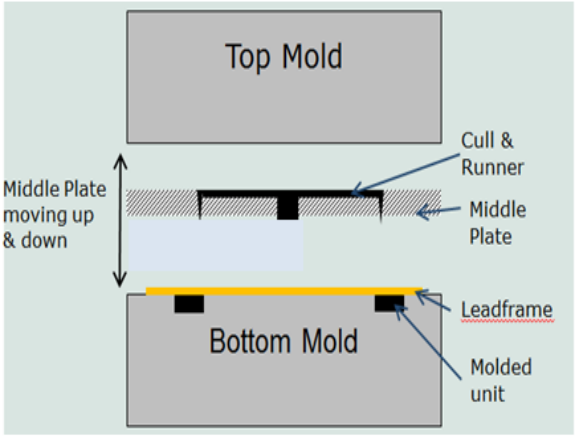


Figure 7. Pin gate mold design

Further investigation found that two individual heater controls are designed into this molding system one on mold housing and the other on mold chase. This leads to complexity of temperature synchronization of two controls: heating and cooling ones [16]. Figure 8 showed the schematic illustration of the heating and control of pin gate molding system. The heater locations and thermocouples and their number can cause the response time to delay for the heaters to turn on or off. All these factors have caused the fluctuation of the temperature of equipment A to wider ranger. Within a molding system, there are four presses. Figure 9 showed temperature variations within a press and between press-to-press.

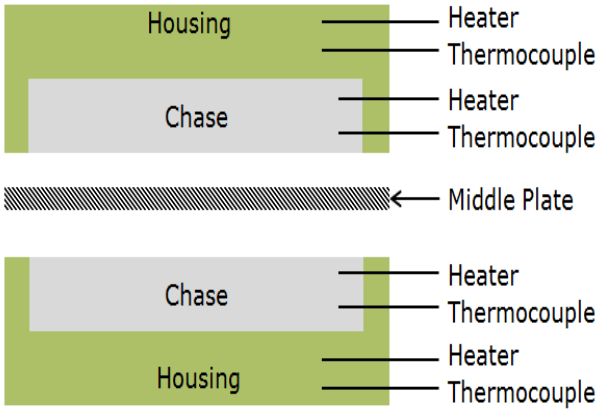


Figure 8. Two individual heaters in one mold press

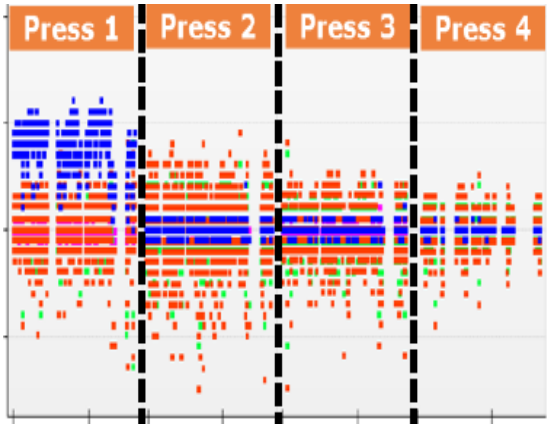


Figure 9. Recorded online temperature variations for four molding presses

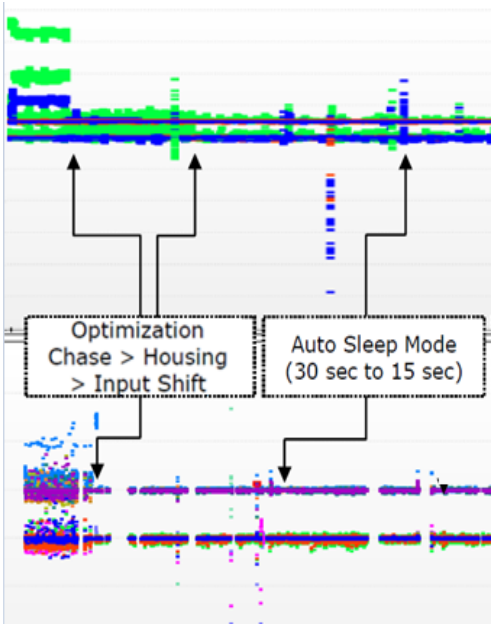


Figure 10. Mold temperature monitoring results after actions taken

After empirical data collection, analysis of the data, and understanding the molding hardware, two key actions are

formulated to overcome the design limitations: a) Single housing heating by disable chase heater to eliminate the temperature control synchronization, and b) Minimizing the waiting time for auto sleep mode to minimize the heat losses at middle plate and cavity. With the two key actions taken, mold temperature has been stabilized as shown in Figure 10.

Figure 10 also showed that sporadically mold temperature can fluctuate on a wide range. Thermocouple wire connector design is the third weakness observed, which can cause unstable temperature. The existing design with thin thermocouple wire attached to TC connector which is easy to spoil if there are any mishandling during install. An enhanced connector design which is more solid and user friendly being introduced into the new molding system. The robust and user friendly connector will not be only replaced by the existing one used in the production floor, but new machine will be introduced with this new connector.

4. Transfer Pressure

In addition to molding temperature, transfer pressure is another key controllable process parameter which can greatly influence the molding quality. In this paper, molding equipment F is used to study the effect of transfer pressure due to its equipment with online real-time pressure monitoring. In this equipment platform, transfer pressure in is monitored under a system called Final Transfer Pressure Control (FTPC). It measures the applied pressure and stops the transfer movement when the required transfer pressure has been reached.

4.1. Process Monitoring Fundamentals

Figure 11 shows a transfer pressure graph in which the transfer position is against the transfer pressure.

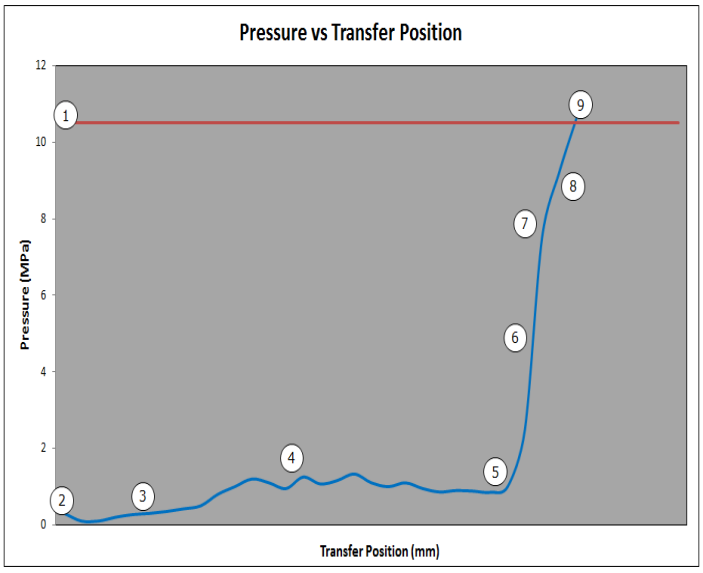


Figure 11. Transfer pressure graph

A description of the event taking place at each number in figure 11 is given below:

1. Red line is transfer pressure set point (i.e. 10.5 MPa).
2. After compound pellets are loaded into the press and they were heat up while the mold closes, the transfer starts moving from its start position which is called pellet load position.
3. There will be a light pressure built up because of the resistance of the plungers in the sleeves and partly melted pellet is pushed up in the sleeve.
4. The compound melts and the transfer moves further up. During the transfer movement, the pressure slightly increases because fluid compound goes through the runners and the narrow gates, which are expected having some resistance.
5. During this transfer process stage, all compound is melted and the pressure built up is caused by the resistance of the compound in the runners, at the gate and in the package.
6. Now, the cavities are completely filled, the air is pushed out through the air vents and the pressure starts to build up in the package. The position from which the line starts rising rapidly equals the cull height.
7. During the fast increase of the pressure, the plunger springs are compressed beyond the spring preload pressure.
8. From this point, the transfer starts compressing the spring.
9. At the final stage, the plunger springs are compressed. The slope of the line depends on the type of spring used in the mold. The transfer move stops when final transfer pressure (required packing pressure) reaches the set point, mold process is completed.

4.2. Transfer Pressure Mechanical System

The transfer pressure mechanical system is dictated by two critical components: 1) Plunger beam and 2) plungers as shown in Figure 12.

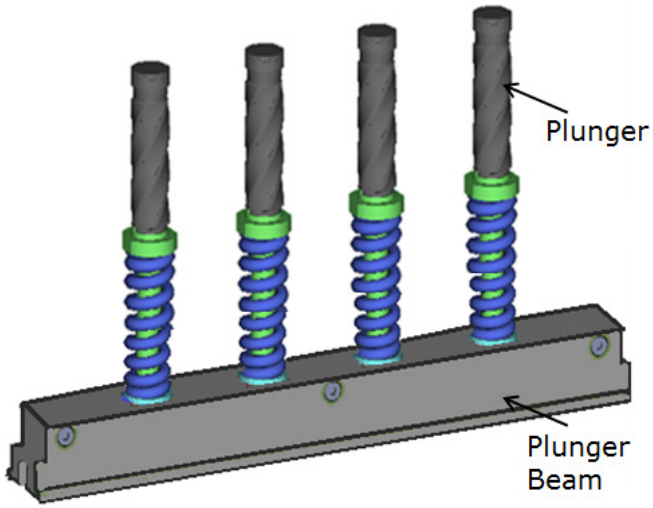


Figure 12. Plunger and plunger beam for molding compound transfer

4.3. Plunger Beam and Plunger

Figure 13 showed the cross section view of an assembled plunger system. The plungers are fitted under a pre-defined tension by means of the compression spring (C). Different spring diameters (recognizable by color type) are used, depending on the required spring force and transfer pressure. The higher part of the plunger (F) is screwed to the lower plunger part (shoulder pin (E) and spring (C)) by means of a long socket screw (D).

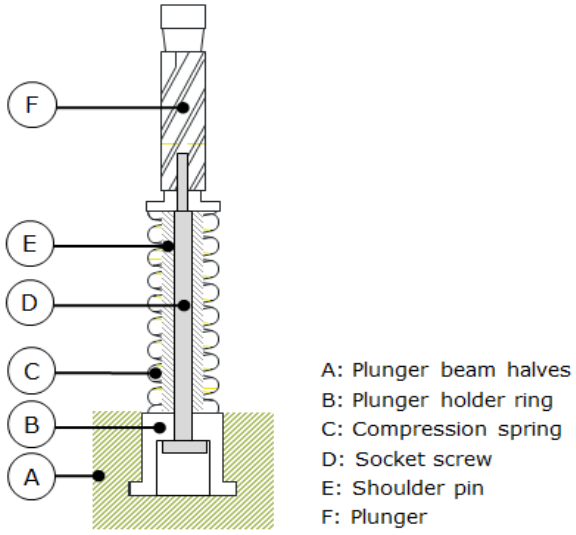


Figure 13. Cross sectional view of plunger beam and plunger

4.4. Transfer movement

The transfer movement consists of four critical pre-set positions. The first position is the starting position which is in compound pellet load position. The second position is the pellet crush position. At this position, the plunger beam waits until the pellet preheat time has finished. The third position is the "Cull Height" (CH) position. Generally, at this point, the plunger cannot move up higher and the cavities are completely filled. But, the transfer movement is not yet applying pressure to epoxy. The end position is the plunger move up to the CH position but the plungers themselves cannot go higher because the cull is blocking this. However, the transfer beam upward movement will cause the cull to compress further by the flexible spring. Figure 14 showed the four critical positions of plunger beam and plunger positions.

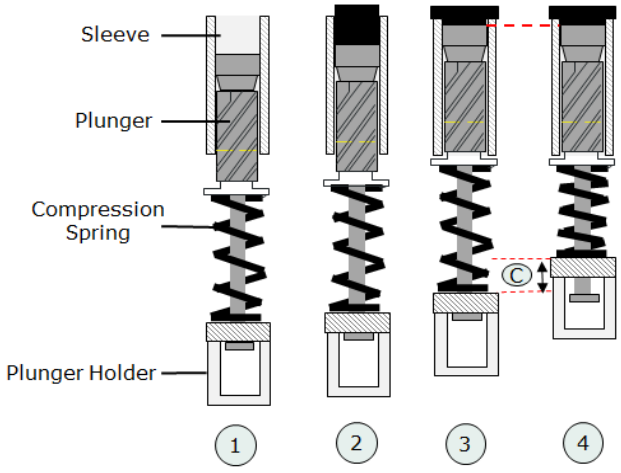


Figure 14. Four critical plunger beam and plunger positions

Cull height setting is one of the crucial parameters that can influence the compactness of the package. However, when the cull height parameter is set lower than the actual cull height, this may deform the product with deform dambar. Therefore, the set end position must be always lower than the cull height. If not, the maximum transfer pressure cannot be attained.

4.5. Transfer Pressure Effects

Transfer end pressure can go higher through the setting of higher plunger position. Different spring type used can deliver different transfer end pressures which purely depend on the cavity designs. When the transfer pressure is too low, this may cause an incomplete fill and air in the package. When the pressure is too high, this may cause die damage or compound leakage. For this paper a case study of plunger beam movement effect on final packing is examined. Figure 15 showed two molding defects due to insufficient transfer pressure.

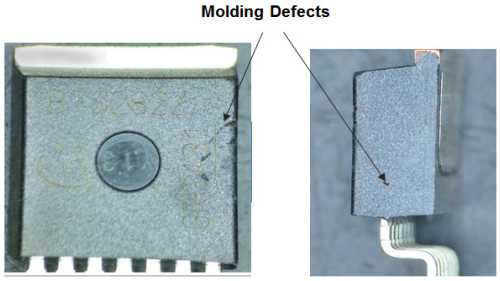


Figure 15. Two modes of molding defect caused by insufficient transfer pressure

The transfer movements and settings in section 4.4 are used as the base for the investigation. The frequent FTPC error triggering at the end filling stage has led to the finding; at the end position the plunger beam upward movement has obstruction. This incomplete final movement had prevented the final compactness of the compound to happen. The root cause of this incomplete plunger beam movement is resulted from fabrication error of plunger holder as shown in Figure 16 below.

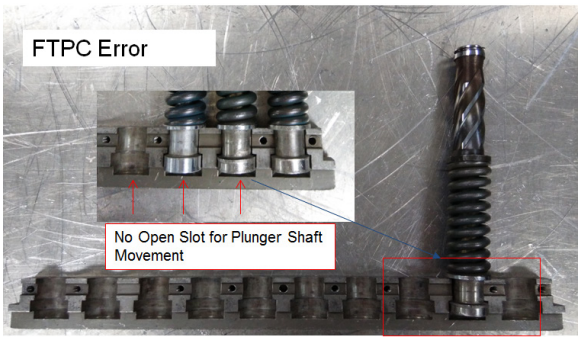


Figure 16. Plunger beam with un-through slot for plunger shaft movement

A new plunger holder with through slot was fabricated and installed into the molding machine as in Figure 17.

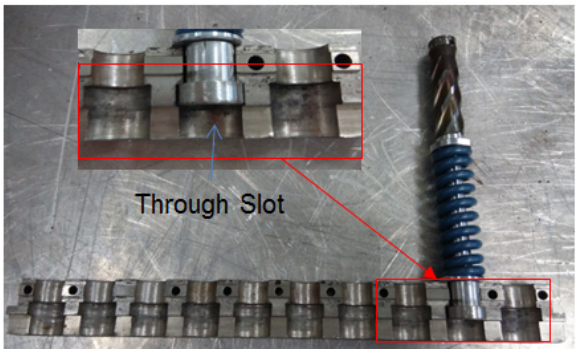


Figure 17. A plunger holder beam with through slot to facilitate free plunger movement

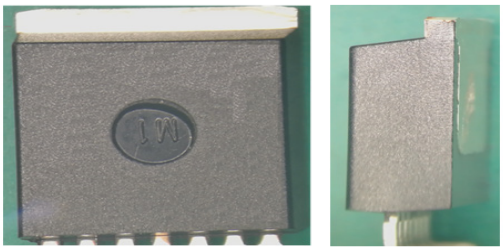


Figure 18. Improvement of molding defects after the corrective action implementation

4.6. Transfer Pressure and Transfer Time Correlation

The effect of transfer time and transfer pressure on short mold and void formation during encapsulation is evaluated in this study. Both short mold and void formation during transfer molding process will affect the long term reliability [17-19] of the package being molded. The tool being used to visually quantity the defect of short mold and void formation is low power scope. Others primary molding process parameters such as top and bottom mold chase temperature, mold compound transfer time are fixed with nominal setting during the evaluation. To minimize the influence of the cleanliness of mold tool on the results, mold die cleaning and conditioning were performed prior to the test. The cleanliness is extended to all the air ventilation area, runner and pot area as well.

The transfer time is defined according to the recommended mold temperature from the melting viscosity curve for compound material A as shown in Figure19. Output response on short molding and void percentage rate are computed and samples are inspected under low power scope by one dedicated inspector. Table 1 and 2 show the Design of Experiments matrix for transfer pressure and transfer time respectively.

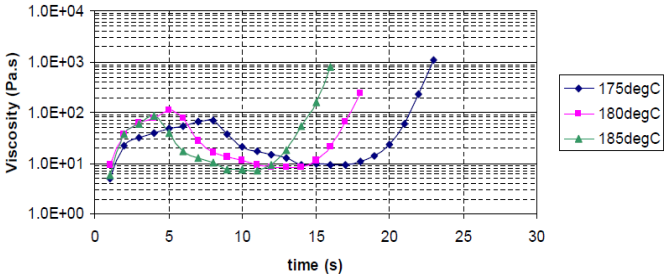


Figure 19. Melting Viscosity chart for compound material A

Table 1. Experiment Matrix of transfer Pressure

DOE Matrix	Mold Temperature	Transfer Time	Transfer Pressure	Remark
1	Fixed	Fixed	Low	Control
2	Fixed	Fixed	High	High Transfer Pressure

Table 2. Experiment Matrix of transfer time

DOE Matrix	Mold Temperature	Transfer Time	Transfer Pressure	Remark
3	Fixed	Medium	Fixed	Control
4	Fixed	Low	Fixed	Shorter transfer time
5	Fixed	High	Fixed	Longer transfer time

High density manufacturing platform is utilized to carry the evaluation. DOE matrix 1 and 2 from Figure 21 show the short molding and void defect inspection results from high density lead frame. The DOE samples are molded with 2 different transfer pressure, low pressure and high pressure. The remaining others primary parameters are all set at nominal values. There is no significant improvement being observed in term of failure rate when transfer pressure increased from low to high with the fixed transfer time [16]. This confirmed that the air trap in the cavity is not due to insufficient compactness during the material transfer to cavity. To gain further understanding, short-shot analysis and mold flow simulation are carried out to further validate the void formation mechanisms and the final air trapped location.

Figure 20 shows the result of shot-shot simulation. It shows the similar behaviour where the air is trapped at end mold flow position and leads to short mold or void formation, which is matching with the current molding defect and location.

DOE matrix 3, 4 and 5 from Figure 21 show the short molding and void inspection result. No significant difference was observed from the matrix 3,4 and 5. According to the viscosity curve, control parameter is the optimum point to achieve the most "liquefy" mold compound flow during the transfer of compound to cavity. The optimum processing parameters, transfer mold is locked at existing parameter for further DOE.

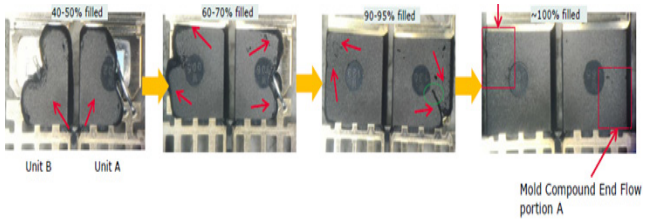


Figure 20. Short-Shot simulation result

DOE Matrix	Mold Temperature	Transfer Time	Transfer Pressure	Result
1	Fixed	Fixed	Low	2F/800
2	Fixed	Fixed	High	2F/800
3	Fixed	Medium	Fixed	2F/800
4	Fixed	Low	Fixed	1F/800
5	Fixed	High	Fixed	2F/800

Figure 21. DOE Matrix result for transfer pressure and transfer time

5. Conclusion

In order to improve the molding processing and reduce the rejection by more than 10%, the following three areas have to be focused: 1) enable only single temperature control instead of dual temperature control to stabilize the temperature. 2) To overcome the challenge of molding process, consideration of internal product structure's design such as wire size, die size, number of wires, and wire orientation which can cause unwanted mold defects. In addition, up-front analysis by adopting numerical tool to understand the existing design and improvement options are proven to be useful. 3) Fix the fabrication error of the plunger holder to overcome the incomplete plunger beam movement which resulted insufficient transfer pressure. On top of that, online process monitoring such as Advance Process Control (APC) is not

only useful to control the process but also a good tool to diagnose the design weakness.

References

[1] K.A.Z. Abidin, K.C. Lee, I. Ibrahim and A. Zianudin, " Problem Analysis at A Semiconductor Company: A Case Study on IC Packages", Journal of Applied Sciences, 11 (2011) 1937-1944.

[2] J. Antony, "Taguchi or classical design of experiments: A Perspective from practitioner" Sensor Rev., 26 (2006) 227-230

[3] M.R. Alam, M.A. Amin, and M.A. Karim, "A Computer-aided Mold Design for Transfer Molding Process in Semiconductor Packaging Industry" ScienceDirect Procedia Manufacturing 21 (2018) 733-740

[4] T.S. Lundstrom and B.R. Gebart. "Influence from process parameters on void formation in resin transfer molding', Polymer Composite Polymer Composites, 15 (1994) 25-33

[5] V.R. Kedari, B.I. Farah, K.-T. Hsiao, "Effects of vacuum pressure, inlet pressure, and mold temperature on the void content, volume fraction of polyester/e-glass fiber composites manufactured with VARTM process", Journal of Composite Materials, 45 (2011) , 26, 2011 2727-2742

[6] T.J.S. Anand, K.Y. Chua, Y.S. Leong, W.K. Lim and M.T. Hng, "Microstructural and Mechanical analysis of Cu and Au interconnect on various bond pads" Current Applied Physics 13 (2013) 1674 – 1683.

[7] M.S. Patel, C.P. McCluskey, and M. Pecht, "Effective decapsulation of copper wire-bonded microelectronic devices for reliability assessment, Microelectronics Reliability 84 (2018) 197-207.

[8] M.S. Krishnanramaswami, G. McCluskey, M. Pecht, "Failure mechanisms in encapsulated copper wire-bonded devices", IEEE 23rd International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA), 2016.

[9] V.B. Willems, "Early fatigue failures in copper wire bonds inside packages with low CTE green mold compounds", 2012 4th Electronic System Integration Technology Conference, 2012, http://dx.doi.org/10.1109/estc.2012. 6542110.

[10] D. Andrews, L. Hill, A. Collins, K.I. Hoo and S. Hunter, " Copper ball bond shear test for two pad aluminum thicknesses, 2014 IEEE 16th Electronics Packaging Technology Conference (EPTC), 2014, http://dx.doi.org/10.1109/eptc.2014.7028415.

[11] S. Manoharan, S. Hunter and P. McCluskey, "Bond pad effects on the shear strength of copper wire bonds, Electronics Packaging Technology Conference (EPTC), 2017 IEEE 19th, IEEE, 2017, December, pp. 1–6.

[12] L.T. Tan, C.H. Lee, Y.Y. Teo and B.H. Lim, "Perfect Molding Challenges and The Limitations", IEEE 19th Electronics Packaging Technology Conference (EPTC).

[13] H. Ardebili and M. Pecht, "Encapsulation Technologies for Electronic Applications", William Andrew, Oxford 2009.

[14] E. Ridengaoqier, R. Fujiki, S. Hatanaka and N. Mishima, "Study on estimation of void ratio of porous concrete using ultrasonic wave velocity", Journal of Structural and Construction Engineering, 83 (2018), 943-951.

[15] G.M.C. Magalhães, G. Lorenzini, M.G. Nardi, S.C. Amico, L.A. Isoldi, L.A.O. Rocha, J.A. Souza and E.D. Dos Santos, "Geometrical evaluation of a resin infusion process by means of constructal design" International Journal of Heat and Technology, 34 (2016) S101-S108.

[16] M. Kubouchi, H. Sembokuya, S. Yamamoto, K. Arai and H. Tsuda, "Decomposition of amine cured epoxy resin by nitric acid for recycling", Journal of Society in Materials Science, Japan. 49 (2000) 488–493.

[17] N.M. Li, D. Das and M. Pecht, "Shelf life evaluation method for electronic and other components using a physics-of-failure (Pof) approach", Machinery Failure Prevention Technology (MFPT) Conference, MFPT, 2017.

[18] N.M. Li and D. Das, "Critical review of US Military environmental stress screening (ESS) handbook", Accelerated Stress Testing & Reliability Conference (ASTR), IEEE 2016, 1-10.

[19] F.P. McCluskey, N.M. Li and E. Mengotti, "Eliminating infant mortality in metallized film capacitors by defect detection", Microelectronics Reliability 54 (2014) 1818–1822.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 492-500 (2018)

www.astesj.com

Special Issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Towards Adoption of Authentication and Authorization in Identity Management and Single Sign On

Ujjwal Joshi[*], Sangwhan Cha, Saeed Esmaili-Sardari

Computer Science, Harrisburg University of Science and Technology, Pennsylvania, USA

ARTICLE INFO

Article history:
Received: 10 September, 2018
Accepted: 04 October, 2018
Online: 25 October, 2018

Keywords:
Authentication
Authorization
Single Sing On
Identity Management

ABSTRACT

Identity and Access Management (IAM) and Single Sign on (SSO) are two security concepts that are related to each other. IAM governs the user access in an organization whereas SSO facilitates the user by authenticating to one centralized application and not having to re-authenticate when trying to access other applications. This paper addresses the different benefits that an IAM and SSO tool can provide to reduce the security risk within an organization. Since, authentication and authorization are one of the major concerns in the cyber security; this paper analyzes common problems that are faced during authentication and authorization. We have also analyzed prior researches that have been done in the IAM and SSO space along with conducting a survey to understand the different issues and benefits of IAM and SSO. From the survey that has been conducted, we have addressed different issues when implementing IAM and SSO solutions along with understanding the architecture, in which these solutions have been deployed into. The surveys conducted have been compared with prior researches done in IAM and SSO space to understand the benefits that the solution provides. The results from the survey have been analyzed to provide the best practices when implementing IAM and SSO solutions along with the benefits provided by the solution.

1. Introduction

Identity Management also commonly known Identity and Access Management (IAM) refers to the set of processes that can be applied to grant right access to the right people within any corporate enterprise environment [1]. Businesses today, implement IAM solutions so that it provides them with the framework that can manage user IDs and passwords along with solving problems related to the challenges associated with managing accesses and permissions of multiple user IDs [2]. IAM solutions also provide different auditing capabilities that enable the managers and higher officials within organization to keep track of the different accesses that their employees have [2]. A terminated employee still having access to the organization resources can cause certain damage to the organization so having an IAM system can reduce the risk associated while manually removing or adding users to different systems. Today, architectures within an organization can be categorized based on whether their applications are deployed in the cloud or within their own network also known as on-premises [1]. This increases the complexity of managing users and granting accesses to them due to increased privacy and security risk [3]. Although IAM systems can remove and create access, the user still needs to enter their user name and password to access the systems that they were provisioned to by the identity management systems. Remembering passwords and usernames can lead to issues where the user can forget their credentials or lose their credentials for certain system, which might fall into the hands of an intruder. Single Sign On (SSO) provides a better functionality for managing authentication and authorizations to applications, by not having the user re-enter the credential every time they need to login [4]. SSO provides a single source of authentication mechanism which provides the user a single platform to enter their credentials to access different applications and also discourages the need to maintain multiple credentials to access different systems [5]. IAM and SSO solutions provide a secure way for businesses to manage and authenticate their users. This research explores the authentication and authorization mechanism that IAM and SSO provides with the view of providing better security while implementing these solutions. The research also explores the different models of architectures that an enterprise environment can have to recommend the best practices that can be followed while implementing IAM and SSO solutions.

[*]Ujjwal Joshi, 696 Fox Ave, Lewisville, TX, USA, 469-451-9885
ujoshi@my.harrisburgu.edu

Security is one of the major concerns that an organization faces today, as risk from cyber-attacks can damage an organization [6]. As business today are growing bigger and faster than anticipated, they are looking for new technologies that can provide a simplified solution for them to manage their user's accesses and resources. A user within an organization needs to have the right set of access to do their work and manually managing user accesses is not a suitable practice for an organization since there can be gaps such as assigning wrong or incorrect accesses [1]. Moreover, manually adding and removing accesses can lead to issues such as forgetting to remove accesses when users leave organization or adding accesses when users join organization.

Since authentication and authorization are one of the major components of IAM and SSO, the need to provide access and manage user authentication are highly in demand. Corporate environments today are complex because they have users and data scattered in different applications and servers. Therefore, there is always a need to manage users efficiently so that the organization is less vulnerable to intrusions and outside attacks. Most organizations today need the security knowledge that can help them to investigate the security solutions that are available. Business always makes mistakes by believing the vendors or service providers rather than understanding their own infrastructure for their security purposes.

Lack of security knowledge and security principles are one of the common problems that organizations face today. By tradition software applications within an organization were supposed to be deployed within the organization boundaries [1]. However, today an organization cannot restrict itself to the traditional views and are deploying applications outside an organization boundary. Any application that is outside the organization boundary is not within the organization trusts zones and organizations do not have complete control over it. Since some applications are within the organization boundary and some are maintained by service providers, the complexity to sync users between application increases. As technologies are rapidly changing today, IAM is gaining a popular momentum as it is one of the key components when managing users on applications deployed within organization boundaries and outside organizations boundaries [3].The problem still lies on how business and organizations can securely manage users and their credentials for applications that are inside the organization and outside the organization boundaries. Since service provider's applications are managed by service provider's resources and application within organization boundaries are managed by organizational resources, manually adding the users and creating credentials can lead to gaps where the user may not have the right access. The complexity also increases when users try to access service provider's applications with different credentials which can lead to forgetting or losing passwords. Such complex situations can be managed by using SSO for authentication and authorization along with IAM for managing user accesses. However, not all the SSO and IAM may be feasible for an organization. Depending of the complexity of the organization infrastructure there is a need to do research which provides guidelines to business so that they know the check and balances that is required when implementing IAM and SSO solutions.

In this paper, we evaluate different criteria that an organization can have from the infrastructure and architecture perspective. The architecture depends on the number of applications that are hosted within and outside the organizational boundaries. The infrastructure is based on the number of employees that an organization has and the different kinds of policies that is applicable to each employee. Then, we categorize the organization based on architecture and infrastructure that the organization has. Depending on the categories that the organization fits into we analyze the different process that an organization can undergo to select the suitable IAM and SSO solutions. We also provide the guidelines on some of the best practices that the organization can follow to implement those solutions. However, this research does not recommend any solution that exists in the market today but provide more of a guideline that the organization needs to take into consideration while implementing IAM and SSO solutions.

The remaining of this paper is organized as follows. Section 2 presents related work. Section 3 describes the proposed approach. The experiments for the proposed approach are described in Section 4. Section 5 provides the concluding remarks.

2. Related Work

Authentication and authorization are one of the mechanisms used in computer security to validate if the user has the right credentials and permission to access a system. However, authenticating and authorization also known as identification, is one of the common problems in computer security [4]. Although many researches have been done to improve the authentication standards, passwords is still common authentication mechanism [4]. Using passwords as an authentication mechanism increases complexity in an enterprise environment since forgetting passwords from time and again can lead to inefficiency while accessing systems. Organizations today are also demanding systems that help them manage users so that end users can manage their private information more efficiently [7]. Identity management is one of the key components of an organization since it plays an important role in authentication and authorization [3]. Identity management refers to the process of managing identities so that each identity has the right set of accesses that is needed for them to do the required work in an enterprise environment [1]. Identity management consists of four major modules which are Authentication, Authorization, User Management and Central User Repository [1]. The authentication module deals with validating user credentials where the user can authenticate themselves by providing username and password or through other different authentication mechanisms such a Kerberos, SAML and so on [1]. The authorization modules provide the functionality of validating a user access to different systems. The user access can range from the group of permissions or roles that a user has to access the system [1]. The user management module provides functionality such as provisioning and de-provisioning of users, life cycle management to manage the user accesses, when the user joins the company till the time the user leaves the company. The central repository module delivers a one stop place for reviewing users and different access they carry along with different systems that IAM is integrated to and reading information from different other managed systems [1].The proper deployment of the four modules of identity

management can result in building robust enterprise environment where processes can be automated and limited manual intervention is required.

An enterprise environment today, is not just restricted to application deployed within a particular network, but also includes various other applications on different vendor systems. Enterprises now are also moving their resources to cloud and are exploring the different cloud models. There are three cloud concepts which are gaining popularity such as Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [8]. The concept of SaaS, PaaS and IaaS has also changed the way different authorization and authentication mechanism worked. The mechanism to store identities across different systems using SaaS, PaaS and IaaS is also known as federated identity management [4]. With cloud-based applications and the need to access applications anywhere at any time has brought in changes in the traditional authentication mechanism and developed different standards that are also known as SSO standards. However, enterprise still use username and passwords as a common authentication mechanism and these can lead to certain problems. In password-based authentication the user passes in a username and password to prove they have the right credentials to access the system. So every time a user needs to access a system they need to enter the password. The use of password-based authentication and classical authorization techniques are described below

- Alpha numeric passwords are used which sometimes can be easy to guess so an unauthorized user may also access the system [4].
- Since different system may have different passwords the user has to memorize all the password or maintain a list somewhere else which can be stolen or lost
- Some system with password-based authentication allows grammars in passwords which provides a way for attacks like dictionary attacks [4]
- Two factor authentications can be a better approach but not all the systems follow two factors authentications.
- Once the user is able to authenticate the next process that happens is authorization which allows the user to access certain resources in the system.
- The authorization techniques for classical system are assigning permissions directly which is difficult to manage [4].
- RBAC (Role Based Access Control) is one of the mechanisms used in various systems to grant access to users for authorization [4].
- The RBAC solves the issues with permissions since permissions are assigned within the roles and roles are assigned to the users [4]. This reduces management risk since they only need to be aware of the roles that the user has.

SSO provides enterprises with a mechanism to authenticate and authorize users with a single password. Once the user logs into one system and authentication happens the users does not need to login again to access other applications [4]. This prevents user from having to remember multiple passwords for different applications. However, since there are many different standards that provide single sign on for users, there is a need to understand and analyze how different single sign on systems works and suits for a particular enterprise environment [4].

Today, there are different SSO and IAM solutions that are available in the market. Organizations need to understand how these solutions work for them based on their architecture and infrastructures design. Some of the commonly used identity management systems that exist today are Sailpoint IdentityIQ, Oracle Identity Manager, RSA lifecycle and governance and so on. SSO tools that exist today are Okta, IdentityNow which have the capability of creating and removing users and also providing single sign on for the enterprise users. Rather than focusing on the tools that exists, it is necessary for the organization to understand the protocols that these tools use and choose the right one that fits their organization. Some of the protocols that are widely use and accepted are OAUTH, OpenID, and SAML etc.

Organizations today have also started a centralized authentication mechanism where the user does not have to remember their entire password and maintain a sticky note or list, attached to their computer [5]. This authentication mechanism when configured properly provides security where users do not have to worry about maintaining sticky note of all their passwords to access different systems [5]. To attain a centralized authentication mechanism different standard can be followed and one such standard is Security Assertion Markup Language (SAML) which uses Extensible Markup Language (XML) based solution for exchanging information between service providers [5]. Another authentication mechanism that can be used is OpenID where an OpenID provider acts as an authentication provider to authenticate relying parties [9]. SSO systems such as SAML, OpenID are the first step in reducing the problem of using different password to log in different systems [4]. Aside from SAML and OpenID another SSO standard that is quite popular is Kerberos authentication. Kerberos provides single sign on through the generation of tickets [10]. These systems are also known as federated authentication standards and are used to authenticate users across different platform [11].SSO protocols like SAML, OAUTH and OpenID replaces traditionalauthentication by using a single source of authentication at the identity provider (IdP) [12]. However, each of these protocols also has their own advantages and disadvantages and based on the literature review there are different ways to attack these single sign on protocols. One of the attacks that is possible which using SAML based single sign on is by constructing a false SAML token [13]. Since SAML uses xml to authenticate the users the attackers can modify the xml to attack the system. One of the major concerns for organization today is how to remediate certain attacks that can occur within the organization. No systems today are free from attacks by attackers despite of the fact whether the organization is using identity management and single sign on solutions. Knowing and having information about different attacks can help organization to plan ahead to prevent those attacks. Some of the vulnerabilities and ways that an attacker can attack an organization involve:

- Injecting malicious endpoints: In this attack the attacker creates a malicious end which is used to force the user to enter their user name and password so that the attacker can retrieve the username and password. Once the credentials are

retrieved the attacker uses the known credential to authenticate themselves [12].

- Redirect Attack: In this attack the attacker learns about the user credentials when the identity provider redirects that user by using wrong redirection code [14]. This is possible since neither OAUTH nor OpenID connect specify how the redirection works [14].
- IdP Mix up Attack: In this attack the attacker confuses the resource provider about the identity provider that the user chooses at the beginning of authorization or authentication process to impersonate as the user to access the data [14].
- State Leak Attack: In this attack the attacker forces the browser to be logged under the attacker name at the resource provider [14]. This attack is also called session swapping since it breaks the session integrity property [14].

One of the challenges that many organizations face today is to understand how single sign on works and the best practices to follow while implementing any single sign on protocols. Since most of the organizations are moving their application to cloud and each application are supported by different vendors the traditional authentication mechanism is not suitable. Also, since not every system is threat proof and analysis need to be done when using different authentication and authorization protocols. The literature and journals available do point out the different standards available for single sign on but do not take initiative on predicting how one relates in terms to other. Also understanding how authorization and authentication works with single sign is one of the major concerns that will be reviewed in this paper. The literature review provides information regarding different SSO standards such as SAML, Kerberos, and OpenID but analyzing how authentication and authorization works with the SSO standards is one of the major aspects that needs to be researched more on.

3. The Proposed Approach

In order to understand the challenges faced by different organizations while implementing IAM and SSO solutions it is necessary to interact, understand and analyze the insights of people working in that platform. Proceeding further, reviewing previous researches done in IAM and SSO provides great insights into the data gathering method and analysis. The research will use qualitative method for data gathering. Qualitative method of data gathering involves individual interviews, observations and research. This research will conduct interviews in phone and through online survey with the IAM and SSO resources that are working with different organizations. The research screening is not only limited to a particular organization within a certain region but will cover different regions and organizations inside United States. Aside from interviews we will also review different researches that have been done prior to this research and analyze the results and outcomes of those researches. The interviews to be conducted will be focused more on organizations that have recently implemented IAM solutions and SSO solutions. The interview methods will also look at organizations that are recently planning to implement some form of SSO and IAM solutions since those interviews can provide more insights into the challenges that the organization had to go through in picking the right solution.

Since interviews are the primary source of data for the proposed research, the interviews will help us to build recommendations and guidelines that can be followed while implementing identity management and single sign on solutions. The interview questions are to be more focused on the steps that were taken before finalizing the IAM and SSO solutions and the challenges faced with those solutions. We tend to use these questions to analyze the different circumstances within the organization prior to building the recommendations. The interviews to be conducted will also have questions that are related to the architecture and infrastructure of the organization and the effectiveness of IAM and SSO solutions to provide enterprise security. The response received from interviews is than compared with the different concepts and researches that have been done prior to this research. Different concepts regarding SSO and IAM are also explored by reviewing journal articles, books and prior researches to build a foundation on the existing solutions. The research will also do a comparative study on the prior researches to fill in the gaps that has not yet been covered. The interviews related to the research have questions related to the impact of SSO and IAM in risk management. The risk management section of the questionnaire will be directed towards understanding how SSO and IAM solutions helped in mitigating risk and vulnerabilities within an enterprise environment.

3.1. Methodology Design

Interviews are the primary source of data for this research topic. The interviews questions are designed to help us understand how IAM and SSO solutions are functioning within an organization. The questions presented are within the understanding of people working in different IAM and SSO solutions or people who have some cyber security knowledge. The research questions are designed to understand the different systems of environment an organization has, different SSO protocols used, examine the IAM and SSO concepts within the cyber security space, number of resources that organization has for IAM and SSO and understand the easiness or efficiency provided by using different IAM and SSO solutions. Visual representation of the data collected and gathered is presented in the experiment results section to analyze and understand the feedback from the survey conducted.

In this paper, we classify the data collected into three different architecture models that an organization can have. The models are based on the different applications and users that the organization has. The models are categorized as small, medium and big and analysis is done projecting the different models and evaluating the differences that each model has in compared to another.

The result obtained from data gathering and reviewing different articles is reviewed to understand the impact and various factors that are involved within an organization. It is expected that the research methodology used above will provide us with the information that is needed to understand the organizations limits and their issues in following the best practices that are required. The result obtained from the research is used to build a set of best practices that is needed for any organization while implementing IAM and SSO solutions in cyber security space. This research

evaluates how IAM and SSO solutions can help to mitigate risk in an enterprise environment when best practices are followed.

3.2. Data Rubrics Classification

The data rubrics classification is used to classify organization based on the users, applications, size of the team that an organization has. The recommended practices to be followed can vary based on each organization architecture and model.

3.2.1. Architecture Design Model

The organization architecture is based on small, medium and big based on the number of users that the organization has

Small
- Number of users less than 10000
- Number of applications less than 50

Medium
- Number of users less than 25000 and more than 10000
- Number of applications less than 150 and more than 50

Big
- Number of users more than 25000
- Number of applications more than 150

Size of IAM and SSO team (Team Classification)

The team classification is based on the number of employees and contractors that are directly and indirectly involved in IAM. Indirectly involvement may include employee and contractors that are aware of the IAM system and need to coordinate with the IAM team for different enhancements that needs to be done.

Small
- Less than 5 employees and Contractors directly involved in IAM and SSO
- Less than 10 employees involved indirectly involved in IAM and SSO

Medium
- More than 5 and less than 10 employees and contractors involved in IAM and SSO
- Less than 20 and more than 10 employees and contractors indirectly involved in IAM and SSO

Big
- More than 10 employees and contractors involved in IAM and SSO
- More than 20 employees indirectly involved in IAM and SSO

Complexities with IAM and SSO solutions

5 The system is reliable
4 The system has issues but has helped organization to function better
3 The system is fairly designed and has space of improvement
2The system is poorly designed but can be improved
1 The system is not reliable and needs to be replaced

3.3. Population and Sampling

The data is collected based on the survey that is conducted and evaluated based on the data rubrics classification mentioned in the previous section. The data analysis process is as described below

- Understand the current architecture of the organizations surveyed and classify them based on the classification rubrics
- Analyze the problems and issues that can be potentially faced by different organization who have implemented IAM and SSO solutions
- Analyze the security issues reduced with IAM and SSO solutions
- Analyze the recommended practices that can be followed with IAM and SSO solutions through survey and prior researches
- Compare and contrast the feedback from the survey and other past researches that has been done
- Build set of recommended practices based on the survey or prior researches

User Population Sampling

The interview process involves 20 different people who are working in IAM and SSO and the analysis is based on their feedback and literature reviews. The user population consists of different people with different background on IAM and SSO.

4. Experiment Results

The experiment results are based on the survey that was conducted with 20 participants that are working in the IAM and SSO security space. The experiment results presented below has been analyzed and compared with prior researches to reflect IAM and SSO practices that are being followed. It also looks at the complexity involved with IAM and SSO solutions and the role they play in mitigating security risk.

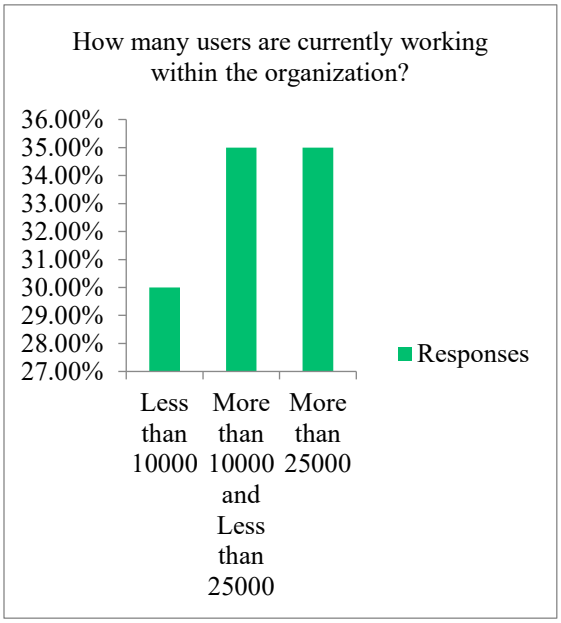


Figure 1: Users currently working within the Organization

4.1. Organization size and Architecture

Based on the survey conducted it can be concluded that most of the organizations that were surveyed had more than 10000 users that was needed to be managed by SSO and IAM solutions. Figure 1 and Figure 2 describes the numbers of users and applications currently managed or needed to be managed by SSO and IAM solutions. 40 percent of the respondents admitted that

the number of applications that has been integrated with IAM and SSO systems is more than 150. Based on this result it can be predicted that the survey data fall under the medium or big tier. From the survey conducted it can also be concluded that most of the organizations engaging with IAM and SSO have a bigger infrastructure and architecture to manage.

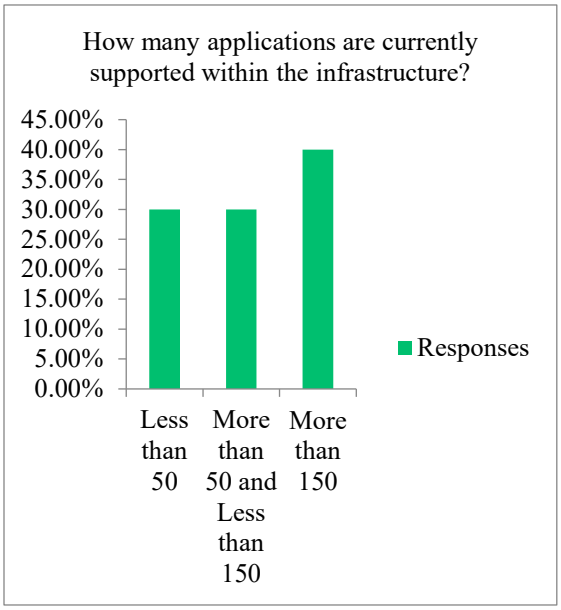


Figure 2: Applications currently supported within the Infrastructure

4.2. Compliance and Auditing Policies

Compliance and Auditing can be described as one of the major reasons that require organization to purchase IAM and SSO solutions. The compliance and auditing policies may differ based on the different domain that the organization is associated with, however the two most popular of those are SOX and HIPAA [15]. Based on the survey gathered majority of the respondents agreed that the organization has some compliance and auditing policies that they need to follow. Figure 3 shows the responses from the participants that were involved in the survey.

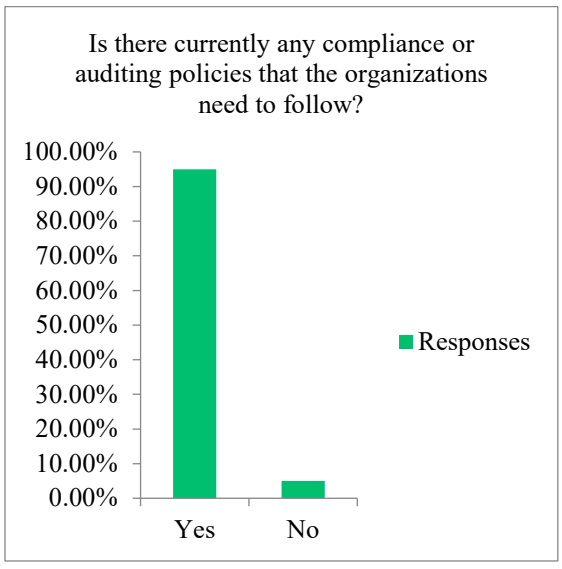


Figure 3: Compliance and Auditing Policies Enforced

4.3. Complexity

Based on Figure 4 most of the respondents responded with already having an IAM and SSO solution in place. Based on this data finding it can be argued that the responses provided to other questions in the survey are more inclined with organizations that have implemented IAM and SSO.

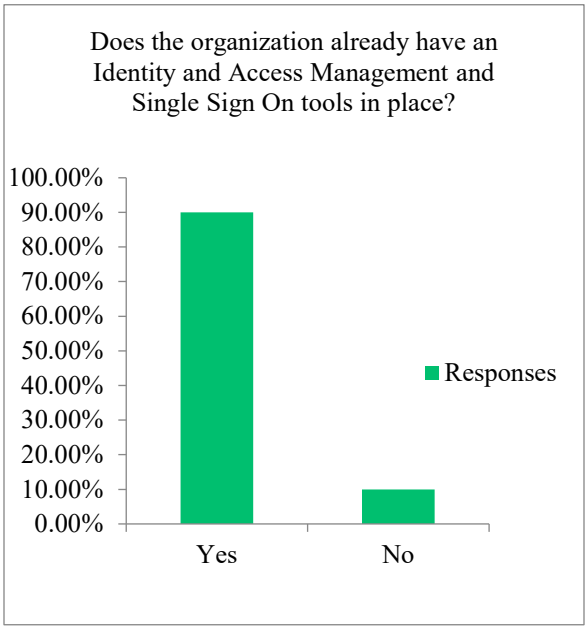


Figure 4: Already having IAM and SSO in place.

The data gathered from the survey also identifies that organizations had reviewed the IAM and SSO solutions before implementing them. The bar graph below shows that most of the respondents responded with the organizations efficiently studying or analyzing the IAM and SSO solutions prior to implementation as shown in Figure 5.

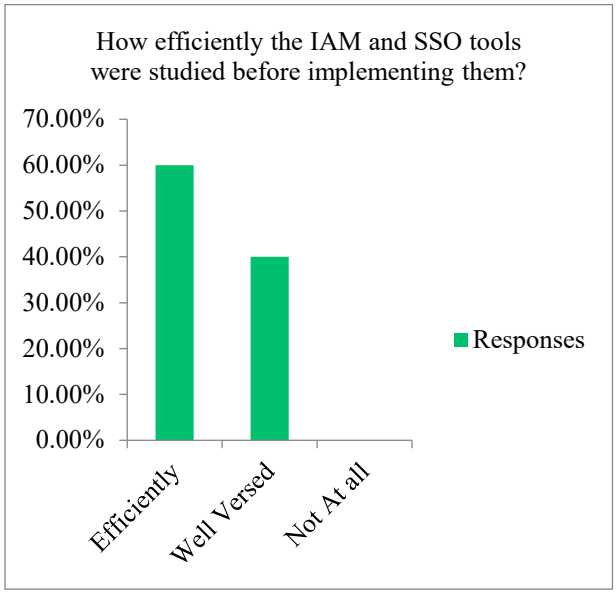


Figure 5: Review of IAM and SSO solutions

Some of the other restriction that was related with the IAM and SSO solutions is shown in the Figure 6.

The Figure 6 defines some of the restriction that is there when using the IAM and SSO tools. Some of the restrictions that are there with the tool are mentioned below from top to bottom according to the data collected are as below:

1. More customization needed with the tool
2. Provisioning capabilities are limited
3. Compliance and Auditing capabilities not as needed by the organization

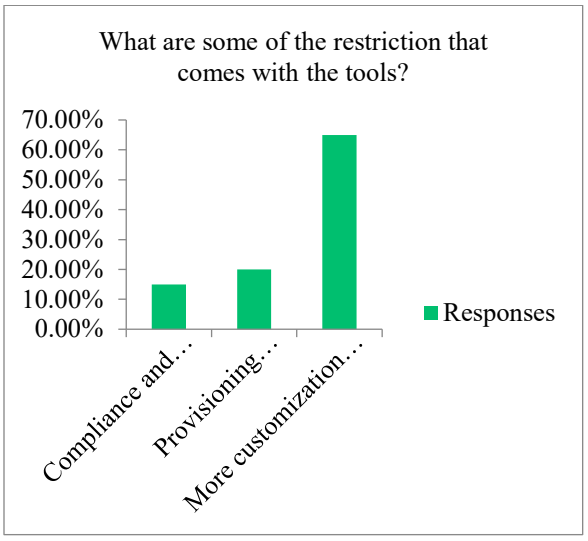


Figure 6: Restrictions with the Tools

4.4. Benefits provided By SSO and IAM solutions

According to the survey conducted and by reviewing the prior researches that have been done in IAM and SSO, we define the different benefits that are provided by Single Sign On and Identity and access management solutions.

4.4.1. Security

One of the major benefits that a SSO and IAM solution are able to provide is the extra layer of security that it adds to the existing infrastructure. IAM solutions ensures that each user in the organization has the right access to do their work so it decreases the risk of employees having more access than that is needed for them. SSO ensures that user do not have to remember their authentication credentials for every application that they need to login in. Removing the necessity of having to remember multiple credentials also reduces the risk of being easily targeted by losing their credentials to the intruders. SSO protocols make it tougher for attackers to easily extract credentials that are caused by server-side vulnerabilities and trying to trick the system to access a particular application which are managed by SSO solutions [16].

4.4.2. Privacy

An SSO solution also helps to maintain privacy within an organization, since employees do not have to re-enter their credentials for every application. Also SSO authentications techniques such as SAML and OAuth depends upon Identity Provider (Idp) and Service Provider (SP) for authentication mechanism the Idp does not have knowledge about the SP that the user is authenticating to which is also known as private browsing [16].

4.4.3. Automation

IAM and SSO solutions are able to provision users to other application without manual intervention or little manual intervention. This functionality helps organization to automate different provisioning, compliance and auditing activities that are needed which helps organizations to grant and remove employees' access faster and with reliability. An organization not using IAM and SSO solutions will need to rely on manual resources for provisioning, compliance and auditing activities which takes more time and effort.

4.4.4. Regulatory Compliance

Organizations within the United States are required to perform compliance activities based on the domain that they operate within [15]. The Sarbanes-Oxley (SOX) and Health Insurance Portability and Accountability (HIPAA) act are among the most popular act that organizations need to abide by. According to Sarbanes-Oxley act of 2002, the SOX act protects the interest of the investors and general public by diminishing fraudulent practices with the enterprise and improving the accuracy of the disclosures. The SOX act applies to all the organizations in US and the organizations are required to comply with the SOX [15]. The HIPAA act which provides data privacy and security within the health and medical domain. One of the major reasons of using IAM solutions at least in USA is also because of the Sarbox or SOX and HIPAA since organizations need to comply with that act [15].

4.4.5. B2B Collaboration

Most of the organizations today do not work alone but rather collaborate with other organizations as partners. This brings in the need for IT systems in one organization to collaborate with other organizations IT systems [17]. Collaboration with other organization needs resources of one organization to be able to access the systems available in other organizations. For this particular reason an employee within a particular intranet of one organization needs to access applications that are in a different intranet. An SSO and IAM tool makes this possible by centralizing their authentication and authorization mechanism and allowing their users login once and be able to access shared resources across multiple organizations [17].

4.4.6. Simple Administration

Having a centralized system makes it possible for enterprises to have a single source where different access that the user has can be defined [17]. The IAM and SSO solutions provide interactive UI which makes it possible to be a one stop shop to view all the users and their accesses. The ease of being able to navigate through different users and revoke/remove certain access that the user should not be having provides organization with greater flexibility of managing their users.

4.5. Challenges Faced by IAM and SSO solutions

The challenges faced by IAM and SSO solutions mentioned below are derived based on the survey that was conducted and prior researches that describe the issues related to SSO and IAM.

Some of the common challenges are as follows:

- Lack of research done on the tool to study whether it meets the requirements for the IAM and SSO solutions
- Complexity of the architecture and applications that needed to be integrated with the IAM and SSO solutions
- Lack of proper documentation of the product and the tool
- Having incident management and change control in place.
- Understanding the vulnerabilities of the tools and how it can impact the infrastructure
- Sufficient and Knowledgeable resources needed to support the tool.

4.6. Best Practices to Follow

After analyzing the survey some of the best practices that can be followed during IAM and SSO implementation are as follows

- The organization should review the tool that they are implementing and understand the complexities associated with the tool
- The organization should have sufficient and trained resources to manage their tool
- The organization should maintain proper documentation for any changes or customization that is done to their IAM and SSO tool
- The organization should have incident management and change control activities in place for any changes that are done to the IAM and SSO tool
- The organization should understand the vulnerabilities that are with the tool and have teams and resources to mitigate the vulnerabilities with the tool
- There should be 24/7 support for monitoring and managing the tool like having an incident response team available

4.7. IAM Data Flow and Architecture

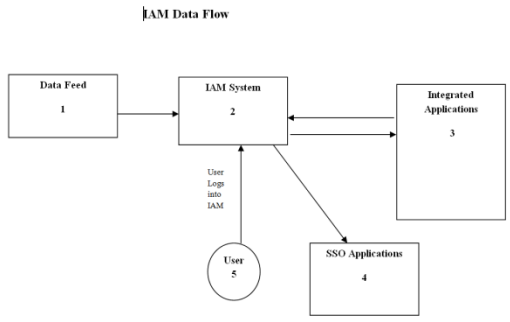


Figure 7: IAM Data Flow

Figure 7 represents the data flow of an IAM system. The bi-directional arrow between 2 and 3 represents the data flow from both the system where one system is the IAM and the other is the applications integrated with the IAM system. The IAM system reads in the user accounts, accesses that each user has in the integrated system but also has the capability to provision new accounts and access associated with users to the integrated system. The data-feed acts as a source of truth for the IAM system from where the IAM system reads in its primary data. The user in the figure represents the user who can log into the IAM system and make specific request for different accesses they need using the IAM system access request framework. If the IAM system supports SSO capabilities the logged in users into the IAM system can access applications without being prompted for username and password. The applications supporting SSO should be integrated into the IAM system to allow SSO functionality to the users. The SSO mechanism is handled by one or more of the SSO protocols that the IAM system is capable of working with such as SAML, OpenID etc.

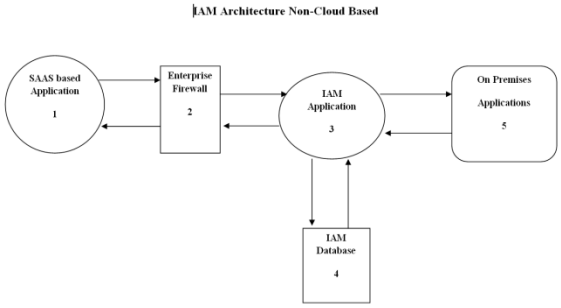


Figure 8: IAM Architecture Non-Cloud Based

Figure 8 represents the non-cloud based architecture of an IAM system. IAM architecture can be categorized as either cloud based or non cloud based. In the non-cloud based architecture the organization is responsible for maintaining the infrastructure needed to maintain the IAM system. The non-cloud based model can be integrated with both the on-premises applications along with cloud based SAAS applications. The architecture consists of an IAM database which is responsible for storing the data that the IAM system reads from integrated applications along with any specific capabilities or data modeling that the IAM system needs to function. On-premises applications are applications that are inside the enterprise network perimeter and the arrows between 3 and 4 represents that the IAM system can read and write data to on-premises application. Since SAAS based applications are not within an enterprise network an enterprise firewall might be present between the SAAS based applications and the IAM system. The firewall should be able to allow request to flow “to and from” SAAS applications to IAM system.

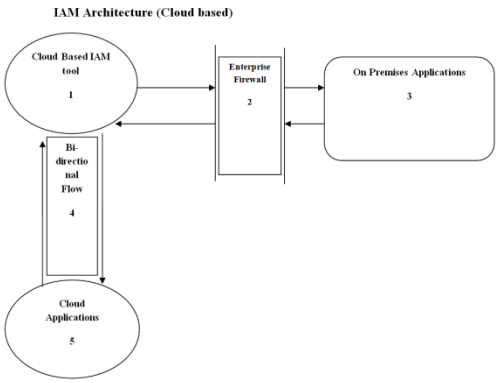


Figure 9: IAM Architecture Cloud Based

Figure 9 represent the cloud based architecture model of an IAM system. In this architecture the IAM system itself is a SAAS application that provides IAM services to the organization. The SAAS based IAM system also provides the functionality to read and write to both on-premises and other cloud based applications.

The arrow between 1,5 and 1,3 represent the flow between IAM system and the integrated applications.

5. Conclusion

In this paper, we analyzed how different IAM and SSO practices are being followed within several organizations. The survey looked at different architectures that the organizations has along with different practices that the organizations followed. In this paper we also analyzed different challenges that the organizations have while implementing IAM and SSO solutions along with the different benefits that an IAM and SSO solution provide. The research also reviewed other research that was done in IAM and SSO to analyze the best practices and comparing it with the survey conducted to understand IAM and SSO practice in organization based on their infrastructure and architecture.

However, we limited our scope to only twenty participants and prior researches to understand the importance of IAM and SSO in the cyber security space. The survey conducted was also limited due to the time and scope of the paper. The participants were only interviewed based on the survey questions and no other possible feedback or inquiry was done aside from the survey questions. Although the research and survey was limited, the responses from the participants helped to analyze the different aspects of IAM and SSO solutions.

Since, our research was only limited to 20 participants, the research scope can be expanded by increasing the number of questions on the survey and the number of participants for the survey. The survey conducted was more focused on understating the importance and benefits of IAM and SSO in the cyber security space, but it did not go in depth to really understand and analyze the IAM practice in different organizations. Since the method of conducting survey was online there was also a limitation on the questions that could be asked and responses that was received. Future work can also look into IAM and SSO practice more in depth from different risk management aspect and cyber security vulnerabilities that can exists. In this research we did not analyze into different cyber security vulnerabilities that can exists with an organization and how IAM and SSO solution can help in reduce each of those vulnerabilities that exists. Future, work can also take into consideration in focusing depth analysis on different SSO protocols that can be used within an organization considering their infrastructure and architecture.

References

[1] Manguic, D. M. (2012). CLOUD IDENTITY AND ACCESS MANAGEMENT–A MODEL PROPOSAL. Accounting and Management Information Systems , 11 (3), pp. 484-500.

[2] Peterson, B. H., Smedegaard, P., Heninger, W. G., & Romney, M. ß. (2008). Managing Multiple Identities. *Journal Of Accountancy* , 1-6.

[3] Nunez, D., & Agudo, I. (2014, 3 6). BlindIdM: A privacy-preserving approach for identity. International Journal Of Information Security

[4] Catuogno, L., & Galdi, C. (2014). Achieving interoperability between federated identity management systems:A case of study. Journal of High Speed Networks 20 , 209-221.

[5] Lewis, D. L., & Lewis, J. E. (2009). Web Single Sign-On Authentication using SAML. IJCSI International Journal of Computer Science Issues , 2.

[6] Andre, T. (2017). Cybersecurity: An Enterprise Risk Issue. Hfm Healthcare Financial Management , pp. 1-6.

[7] Tormo, G. D., Millán, G. L., & Pérez, G. M. (2012, 12 27). Definition of an advanced identity management infrastructure. pp. 173-200.

[8] Alston, A. (n.d.). Attribute-based Encryption for Attribute-based Authentication, Authorization, Storage, and Transmission in Distributed Storage Systems. pp. 1-20.

[9] Hsu, F., Chen, H., & Machiraju, S. (2011). WebCallerID: Leveraging cellular networks for Web authentication. Journal of Computer Security , 862-899.

[10] Pérez-Méndez, A., Pereñíguez-García, F., Marín-López, R., & López-Millán, G. (2012). A cross-layer SSO solution for federating access to kerberized services in the eduroam/DAMe network. International Journal of Information Security , 365-388.

[11] Shitamichi, T., & Sasaki, R. (2014). TECHNOLOGY OF FEDERATED IDENTITY AND SECURE LOGGINGS IN CLOUD COMPUTING. International Journal of Electronic Commerce Studies , 5, 39-62.

[12] Mainka, Christian; Mladenov, Vladislav; Schwenk, Jörg. (n.d.). On the security of modern Single Sign-On Protocols –Second-Order Vulnerabilities in OpenID Connect.

[13] Krawczyk, P. Secure SAML validation to prevent XML signature wrapping attacks.

[14] Fett, D., Küsters, R., & Schmitz, G. (2016). A Comprehensive Formal Security Analysis of OAuth 2.0.

[15] Heino, E. (2011). Evaluating financial benefits of an identity management solution CASE Logica. Retrieved July 9, 2018, from http://epub.lib.aalto.fi/en/ethesis/pdf/12500/hse_ethesis_12500.pdf

[16] Alaca, F., & Oorschot., P. C. (2018). Comparative Analysis and Framework Evaluating Web Single Sign-On Systems.

[17] Bazaz, T., & Khalique, A. (2016). A Review on Single Sign on Enabling Technologies and Protocols. International Journal of Computer Applications, 15 (11), 18-25.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 3, No. 5, 501-509 (2018)

www.astesj.com

Special issue on Multidisciplinary Sciences and Engineering

ASTESJ

ISSN: 2415-6698

Emotional Impact of Suicide on Active Witnesses: Predicting with Machine Learning

Richard Osei Agjei[1], Emmanuel Awuni Kolog[2,3,*], Daniel Dei[4], Juliet Yayra Tengey[5]

[1]*Texila America University, Department of Public Health, India*

[2]*University of Ghana, Department of Operations and Management Information Systems, LG 78, Ghana*

[3]*University of Eastern Finland, Department of Computer Science, School of Computing, FI-8001, Finland*

[4]*Valley View University, School of Graduate Studies, Ghana*

[5]*University of South Africa, Department of Psychology, South Africa*

ARTICLE INFO

Article history:
Received: 12 September, 2018
Accepted: 20 October, 2018
Online: 26 October, 2018

Keywords:
Emotion
Machine Learning
Suicide
Witness-based account

ABSTRACT

Predicting the impact of suicide on incidental witnesses at an early stage helps to avert the possible side effect. When suicide is committed in public, incidental observers are left to grapple with it. In many cases, these incidental witnesses tend to experience the emotional side effect with time. In this study, we employed a Machine learning algorithms to predict the impact of suicide and suicidal attempt on incidental witnesses. This prediction was based on the accounts of suicide given by selected participants who have witnessed the act. The accounts, which was pre-processed into a corpus, were manually annotated with predefined emotion categories. While sadness emerged as the most salient emotional impact on the witnesses, fear was found as the lowest of the emotional impact on the witnesses. However, the machine learning prediction yielded highest in predicting depression with insignificant variations in the other emotional categories. This nonetheless shows that people who have witnessed suicide or suicidal attempts are inherently affected by some form of emotions that may require urgent attention to alleviate. By evaluating the performance of the Machine learning algorithms, the Support Vector Machine was superior, in terms its prediction, then the Multinomial Naïve Bayes algorithm. The outcome of the study contributes to the pool of research that sought to advocate the use of Machine Learning for predicting social phenomenon.

1. Introduction

Suicide is the termination of an individual's own life purposefully due to several factors ranging from work pressure, family pressure, peer pressure, financial pressure, academic pressure, relationship or family disharmony, social pressure, alcoholism among other known and unknown circumstances. These behaviours have become a global phenomenon where a considerable research has proven its association with cognitive and affect disorder. In recent times, the problem of suicide has become a common occurrence in both developed and under-developed countries leading to psychosocial effect and evocation of negative emotions such anger, sadness, depression, fear, worry and anxiety. Among the global violent deaths, suicide accounts for 50% for men and 71% for women [1]. The impact of suicide on second and third parties in the society has become a social issue that requires utmost attention. Globally, the female and male ratio for suicide is 1: 4. This ratio amounts to the seventh leading cause of death among men [2]. However, in developed countries, suicide ratio of women to men is 1:3 while low and middle-income countries have the ratio of 1: 5.

Suicidal behaviour is any action that could cause a person to die. Before committing suicide, people exhibit suicidal behaviour that could be averted if the behaviour is detected early enough. Individual's with suicidal behaviour possess some kind of physical, psychological and economical threat to their families, friends and the society at large due to the rippling effect that

*Dr. Emmanuel Awuni Kolog, Box LG 78, +233248271443 & eakolog@ug.edu.gh

suicide or suicidal behaviour pose. Generally, research has revealed that suicide rate is predominant in the age of 70 years and above while the second leading cause of death is among those in the age range of 15 to 29 years [1]. People commit suicide in several ways. Notable ways are Jumping from a height, drowning, going under a train, hanging, use of firearms, poisoning, self-cutting, self-immolation, and vehicular suicide, [3,4,5,6,7,8,9]. Adequate knowledge of these behaviours can inform several interventional methods and techniques.

When suicide is committed, the emotional impact of the act on witnesses is often overlooked. When the act is committed in public, incidental observers are left to grapple with it. In many cases, these incidental witnesses tend to experience the emotional side effect, which in some cases result in more severe damage. Predicting the impact of suicide of witnesses at an early stage helps to avert the possible side effect. With this in mind, we aimed at, from the point of view of using computational Machine learning techniques, tracking the emotional impact of suicide or suicidal attempt on witnesses from their accounts on witnessing suicides or attempts on suicide. The accounts were analysed with Machine learning classification techniques.

Many researchers have used machine learning applications for the detection of emotions and sentiments in text from diverse domains. In the health sector for instance, where clinical situations or assessment appear ambiguous, machine learning can be employed to validate, ascertain and strengthen clinical prediction [10, 11]. Machine learning can be applied to clinical data and/or data on the measurable indicator of the severity of suicidal behaviour. Thus, machine learning used in conjunction with clinical data or biomarkers will help reduce the error margin in suicide risk and suicide predictions. In the arena of psychology, the use of machine learning is used predominantly to predict suicide risk among patients with mood and depression disorders [10].

2. Theoretical Framework

In this section, these researchers discuss related literature and theoretical foundation of emotions and their impact on suicide. Additionally, the role of using machine learning to predict suicide is discussed in this section.

2.1. Emotion and Suicide behaviour

Humans are a highly social species [12], and that imply that emotion prioritize response to threats and opportunities in an environment in which individuals seek to survive. When friends and family relations witness suicidal behaviours in the environment, where they live, it poses psychosocial effect and negative urgency on their emotional behaviour.

Emotion is considered to be inherently social [13] which implies that the environment and the interacting partners are key players in the evocation of positive and negative emotions. Emotions can trigger heuristic thought or swift action. It is hypothesized that emotions will bring a direct effect on quality of relationship [14], but will also moderate the relationship between its components. Thus emotions exert causal effects on the quality of our relationships [15]. Emotions represent key underpinnings for interpreting not only human decision making but also much of human behaviour as a whole. Since affective researchers have given much attention to the impact of the environment on individual's emotions [16, 17, 18], it is important to consider the impact of suicide on people who incidentally witness suicide. It is therefore crucial to understand how human emotions are affected by constant interaction with their immediate families, friends and the society as a whole [19]. The social functions and impact of society on individual's emotion have seen a considerable research, yet the general agreement among affective researchers as to whether they represent emotion is a matter of constant debate [19].

Decision serves as a conduit for both positive and negative emotion while decision making is largely influenced by emotions [21, 22, 23]. Emotion can influence suicidal thought and the depth of information which can trigger physiological, behavioural, experimental and communicational responses that enable an individual to take action quickly based on the encountered problem [23]. There is a connection between general mood and the emotional situation and happenings in an individual's environment [24]. When individual witness the incident of suicide, the emotions that are evoked are incidental emotions, which often produce influences that are unwanted and unconscious.

Suicide is a complex neurobiological phenomenon, and there is great variability in the validity of clinical assessment tools for predicting suicidal risk. Despite reaction of uncertainty surrounding witnesses' accounts of events, recent scholarship on violence has highlighted the credibility of testimonies by witnesses. Exposure to violent behaviour has the same traumatic effects on witnesses as it does on victims [25]. Like violence of all forms, witnessing suicidal behaviours affect witnesses. Extant studies have revealed that there is a high chance that witness of suicidal behaviours or suicidal attempts can develop maladaptive lifestyles [26]. In some situations, witnesses show symptoms of withdrawal in addition to maladaptibility [27]. When individuals witness suicidal behaviour the rippling effect is the retardation of the normal mental functioning of witnesses [28]. Several research techniques have been applied in determining the effect of suicide on incidental witnesses, notable is from the point of view of qualitative approach. To add is that, less studies have considered the use of predictive techniques, such as machine learning, in this domain of research. This partly informed the reason for this work, in order to ascertain if the use of machine learning can predict the emotional effect of suicide on incidental witnesses of suicide.

Considerable studies have reported that three out of every five suicide deaths yielded no clue of the suicidal communication, threats, ideation/suicide attempts or any life time history of suicidality [29, 30]. More so, there are several illnesses that precipitate suicidal ideation among patients, and these illnesses differ from patients to patients which makes it more difficult to identify a specific biomarker(s) to predict suicide risk [31]. The ability for health professionals and providers to predict an individual's suicide risk with high certainty is limited by lack of clinical prediction rules [31]. The concept of suicidal intent/behaviour in connection with emotional dysregulation is ambiguous and contradictory while examining several studies [32, 33, 34].

Several theoretical models have been proposed in literature to provide strong theoretical and factual information that concerns the progression of puzzling behavioural and emotional outcomes. Notable of these theories of suicidal behaviour is the Interpersonal

Theory of Suicide [35, 36, 37]. The Interpersonal theory of suicide was earlier suggested by Joiner (2005), and later elaborated by Van-Orden *et al.* [36]. According to this theory, "the most dangerous form of suicidal desire is caused by the simultaneous presence of two interpersonal constructs—thwarted belongingness and perceived burdensomeness" [36: 1]. The theory further elaborates that the capability to engage in suicidal behaviour is separated from the desire to engage in suicidal behaviour. The Interpersonal Theory of suicide presents a workable construction to further improve the quality of our understanding of suicidal behaviours. This theory has proven to be one of the best lenses through which suicidal behaviour can be understood [38]. It explains the suicidal intent or ideation-death gap and provide more detail understanding of suicidal behaviour in general. The concept reveals that the interplay between perceived burdensomeness and thwarted belongingness brings about the need for an individual to die by suicide. The theory suggests that the riskiest form of suicidal desire is introduced and provoked by the interplay between perceived burdensomeness and thwarted belongingness. In addition, Christensen *et al.* [39] found that the interplay between perceived burdensomeness and the thwarted belongingness predicts suicidal ideation. It is interesting to note that in their findings interpersonal theory of suicide explained more variance than the epidemiological models using demographic factors, including depression, anxiety, traumatic and stressful events as predictors [39].

2.2. Effect of Negative Emotions on Suicide

Lerner *et al.* [40] view emotion as a "coordinated reactions to survival-relevant event, including cognitive and biological changes, facial and/or bodily expressions, subjective feelings, and action tendencies such as approach or withdrawal". Suicide has a close association with the influence of negative emotions. While emotion is considered as one of the factors to influence suicide, Loyo *et al.* [60] have reported failures in interference control, cognitive rigidity, and alterations in verbal fluidity and decision-making as other influences of suicide or suicidal attempts.

The emotional aspect of the Interpersonal Theory of suicide is important, in that, the three component of the theory are dangerous for any one. This danger is "significantly amplified in individuals with high level of negative urgency, as they may be more reactive to less severe environmental stressors and more highly motivated to respond to negative behaviours capable of impacting suicidal risk" [41]. Depression, anxiety and trauma comes with negative urgency. This tendency to act rashly in an attempt to reduce feelings of negative Affect: unspecified feelings; the superordinate umbrella of constructs involving emotion, mood, and emotion-related traits [42]. Individuals with high levels of negative urgency might experience quicker, more sever increase in suicidal desire as a result of the devastating nature of their negative emotions [41]. Anestis *et al.* [41] argued that negative urgency does not only impact the desire for suicide, but also intensify the development of the acquired capability.

Problems associated with emotions are classified as a significant risk factor for suicide behaviour. Lees and Stimpson [43] explain that the concept of emotion, in reference to suicide, is often neglected due to the misunderstanding of emotional behaviours. Theories that bother on the state of the mind, emotional states and actions of suicidal person assimilated affect by theorising suicidality as self-harm or a form of hostility to self. Researchers have found that people participate in deadly self-harm (suicide) as a way of avoiding undesirable emotions and prevent unwanted pain. For instance, extant study suggests that being conscious of one's shortfalls leads to higher destructive emotion and for that matter, individuals cultivate a longing to end this consciousness [44]. With this effect, individuals try to lessen this consciousness and control emotion thus resulting in irrationality and worthlessness [44]. This state of reduced consciousness and rejection of emotion makes extreme actions, such as suicide, a suitable way to escape from one's self and the world.

Additionally, flashbacks, trauma, anxiety and fear of those who witnessed, experienced or experiencing suicidality is connected with problems in controlling emotions and heightened rates of suicide. Though Thomas [45] maintains that individuals with unstable emotions have inadequate knowledge of their personality or health status, thereby engage in suicidal behaviour as they recognised as the only means of coping. Research investigating emotion susceptibility to suicidal behaviour have shown that there is no single pathway or pattern to follow or track individuals who are the first time suicide attempters. This however suggests that the non-rejection of emotional reactions and poor access to professional emotional treatments could be the major means to foretell modern behaviours that are connected with suicide [46].

Anestis *et al.* [41] maintain that research examining the role of emotion regulation in Interpersonal Psychology Theory of Suicide has been inadequate. This is because the main focus of suicidal behaviour is lack of emotion regulation rather than behaviour. The researchers further reiterated that since emotions affect a number of psychiatric disorders with high suicide rates, it would be important to examine its relationship to suicide behaviour. Fredrickson [47] offers some theoretical and empirical support by indicating that individuals undergoing positive affect show flexible, imaginative, resourceful thought and positive emotions get rid of persistent negative emotions.

3. Research Methodology

This section elaborates on the methodological process of achieving the research objectives. The context of this study and the data collection procedure are as well outlined in this section. Pre-processing of the subjective responses of the participants is coded according to a pre-defined emotion category, which are presented in this section. Given the recent interest in advancing inter-disciplinary research in ICT, these researchers have further analysed the collected data from the point of view of using machine learning classification techniques to predict the emotional impact of suicide on witnesses based on their accounts of suicides or suicide attempts.

3.1. Data collection

Data for this study was collected in Ghana. Nine hundred and seventy-eight (N = 978) instances of emotional antecedents, regarding witness of suicides or suicidal attempts, were collected from randomly selected participants, drawing mainly from the health and other institutions of higher education. The basic rationale for the diverse sources of the data was to help understand the impact that suicide or suicide attempts have on people. The

collection of the data was mainly through digitally-developed questionnaires. Based on the objectives of this study, the selected participants, for this study, were tasked to respond to the questions subjectively, asking for their experience on suicide or suicidal attempts by themselves or others. The questions were purely an open-ended one, except for their demographic information. The open-ended responses from the participants were developed into a corpus that required further pre-processing for easy coding or annotation for the prediction.

3.2. Ethical Considerations

Given that human emotions are a sensitive phenomenon, we took into account the ethical considerations of the selected institutions and participants who responded to the questionnaires. An official letter requesting for the right to collect data was forwarded to the ministry of education and the selected institutions. The major content of the letter was to seek ethical clearance before the data was collected. This process was in compliance with the Ministry's policy of data collection.

What is more, an informed consent was obtained from each of the participants. These researchers took turn to educate the participants on the importance of this study and why their data is relevant for this study. The participants were given the opportunity to opt out at any point should they find any ethical breach or discrepancies in the process. Consequent to this, some of the participants opted out without recourse to their action. Nonetheless, the participants were assured that their data or responses will be treated confidentially and that only group findings will be reported, which will not have a direct connection to individual participants.

3.3. Pre-processing

These authors developed basic emotional themes (categories) that align Plutchik's eight basic emotions [59]. The development of the emotion categories was based on the manual analysis of the sample text instances. In line with Plutchik's eight basic emotions, however, we sampled 10% of the corpus and analysed to determine the common emotional themes. In the end, we deduced seven basic emotions which were found as the various emotions that the participants expressed in the sampled data. These emotion categories are *worry, anger, fear, sadness, depression, surprise* and *no emotion.* The corpus was pre-processed for the selected coders to code each instances of the corpus with the predefined emotion categories.

Three human coders with background in human psychology were selected and trained to label the instances of the corpus with the predefined emotion categories. The reason for the selection of the three coders is backed by Kolog *et al.* [49]. Kolog *et al.* [49] revealed that the emotional state of text coders influences their perception when analysing emotions in text, especially when the coders are given limited period to carry out such task. What is more, the coding process is time consuming and the outcome may not be satisfactory which could result in bad decisions on the side of the coders.

Furthermore, these researchers are of the view that people are more susceptible to the influence of their environment which has the potential to influence their perception of emotions. For example, one's domestic challenge, at a point, could influence one's decisions while analysing other people's emotional behaviour. This could, nonetheless, be attributed to the subjective nature of emotions when especially expressed in text. For this reason, the selected coders were given ample and free time to report their findings. Although, some of the instances of the data could yield or contain more than one emotion category, the coders were asked to choose the emotion category based on the most dominant emotional keywords. Below are two instances of the coded text instances of sample corpus:

> *"After hearing that I felt very sad and wondered why one will take his or her life no matter what one is going through. I feel the devil has gotten it way too much into our society because I know there is a spiritual backing to suicide which is taking much control over many especially the youth*" ***[Surprise].***

> *"Hmmm! it very sad to hear that a fellow age mate has taken their own live. Sometimes I ask what at all they could be going through that they can share with anyone but only thought of suicide. I got depressed for couple of days before coming back to normal. Hmm only God knows Best"* ***[Depression].***

3.4. Machine Prediction

Having labelled the text instances by the selected coders, we further employed a machine learning technique to classify the corpus according to the pre-defined emotion categories. With this, we used WEKA machine learning software [50] for the prediction (classification) of the emotional impact of suicide on the witnesses. WEKA contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization [50]. It is well-suited for developing new machine learning schemes. WEKA was developed by Machine Learning Group at the University of Waikato in the New Zealand [50]. WEKA is a collection of machine learning algorithms for data mining tasks. WEKA machine learning tool was adopted because of its versatility to analyse unstructured text. The tool is a widely known data mining tool that has generally been accepted in this research domain. Additionally, WEKA makes available source code for some of the packages, such as the classifiers, for importation into personal project. This makes it flexible to makes changes to achieve the desired outcome.

Two of the Supervised machine learning classifiers in WEKA were used in this study - Support Vector Machine (SVM) and Multinomial Naïve-Bayes (MNB) classifiers. The choice to use these classifiers is rooted in the fact that considerable studies have found these classifiers efficient and robust for text classification [51, 52]. In WEKA package, we employed 10-folds cross validation for training and predicting of the emotions in the text. This was carried out after the corpus had been tokenised and the features extracted for the classifiers. With this in mind, the feature words were extracted as bag-of-words for the classifiers to learn and create a model for the prediction. The 10-folds cross validation means the coded corpus was randomly partitioned into 10 equal sub-sample sizes. This suggests that the cross-validation process is repeated 10 times (the folds), with each of the 10 subsamples used exactly once as the validation data. In the subsequent paragraphs, these authors explained the various classifiers used in this study. Figure 1 illustrates the various stages in the prediction of the emotions from the witness-based accounts using supervised machine learning.

To ascertain the efficacy of the machine prediction, we evaluated the performance of the classifiers against the human

coded corpus. With this in mind, we computed for the recall, precision, F-measure and the accuracy of the various classifiers. While there are other evaluation measures, precision, recall and accuracy have widely been used in the evaluation of text classifiers. Recall is the proportion of the labelled instances of the gold standard that are identified and extracted by a classifier while the fraction of the automatically extracted data that was labelled correctly as the gold standard by the classifier is Precision. The F-measure, as referred to as the F-score, is the harmonic mean of the recall and precision. The Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations [53].

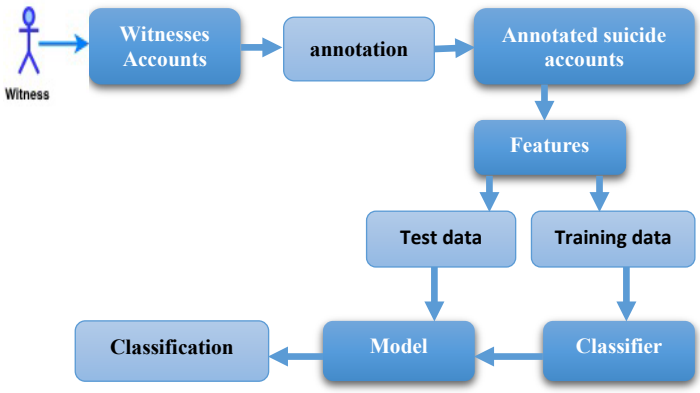


Figure 1. Conceptual classification process

3.4.1. Support Vector Machine

SVM is a supervised machine learning algorithm which can be used for both regression and classification. SVM is a discriminative algorithm in a sense that it is defined by constructing a hyperplane or a set of hyperplanes in high dimensional vector space. Before SVM performs classification task, feature words from the text instances is expected to be extracted through several proposed approaches such as bag-of-words, part of speech tagging among other new advances. After constructing the hyperplane according to how the data is labelled, a model is created for predicting unseen data. Intuitively, "a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general, the larger the margin, the lower the generalisation error of the classifier" [51]. In this work, we used Sequential Minimal Optimisation for SVM in WEKA [50]. Research has indicated that SVM performs more accurately if the training data is very large [54]. Figure 2 shows a typical separation of data points into two categories, thus binary classification. This can, for example, be the detection and classification of input data into positive and negative polarity.

Detecting and separating the data points in a high dimensional vector space is transform by a *kernel*. The kernel in SVM is used to recognise patterns in the data points. So far four different kernels are available in SVM: *linear kernel, polynomial kernel, Gaussian basis radial basic function* (RBF) kernel and *sigmoid kernel*. However, in this study, all the various kernels were used to determine which of the them would yield the best performance. In the end, the linear kernel performed better, of which the results are reported in this work.

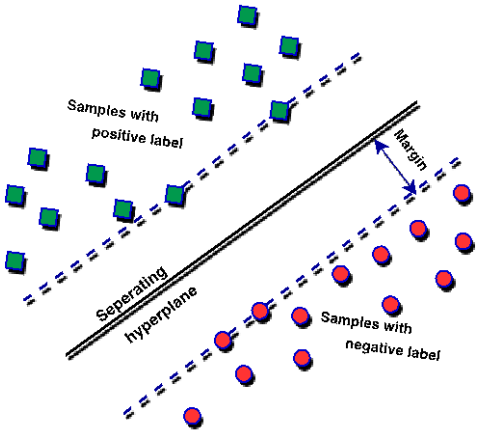


Figure 2. Maximum margin and separation of the hyperplane in SVM (Adapted from Kolog, 2017)

3.4.2. Multinomial Naïve-Bayes

Unlike the SVM, NB is a text classifier that is based on the Bayesian theorem with strong and independence assumption in the model rather than a particular distribution of each feature. A Naive Bayes model assumes that each of the features it uses are conditionally independent of one another given some class. Naïve-Bayes is actually suited when the input data is in high dimensional vector space. What is however different is that NB can classify text instance more accurately even if the training data is small.

4. Results and Discussion

This section reports the findings arising from the analysis and classification of emotions in the witnesses' accounts. This considers the demographic information of the participants, and the emotional impact of the participants from the point of view of their accounts on suicidal antecedents. The results are accompanied with the discussion and backed by literature.

4.1. Participants demography

Of the 978 instances of the text corpus, 49% (n = 483) of them were reported by male participants while the remaining 51% (n = 495) were reported by the female participants. As research has proven that *age* is an important moderating factor in determining suicidal intent or behaviour [55], we considered the age brackets of the participants in this study. With this in mind, 4% (n = 38) of the participants were in the age range of 4 to 24, 49% of the participants were in the range of 25 to 44, 47 % of them were in the range of 45 to 64 and none of the participants were above the age of 65years. Figure 3 shows the gender and age representation of the participants in this study.

Having established the gender and age of the participants, we further prod the participants regarding the demographic information of the supposed people they have witnessed committing or attempting suicide. Although this was not the central focus of this research, the demographic information is equally vital in ascertaining the calibre of people who have committed or attempted suicide. From the responses, we deduced that 43% (n = 419) were female victims of suicide as reported by the participants while the remaining 57% (n = 559) were male victims. This is consistent with extant findings that suicide is

prevalent among men than women. For example, the American foundation prevention for suicide have reported in 2018, that men die by suicide 3.53 times more than women [56]. Additionally, in 2017, WHO reported 1, 129 of suicide cases in men as against 729 suicide cases in female [57]. In Ghana, WHO have reported that 1,500 Ghanaians commit suicide every year where men record the highest number of suicides [57]. However, female has the highest number of attempted suicides and broken hearts in the various psychiatric hospitals in Ghana [57]. Conversely, the participants reported higher number of victims of suicide or suicidal attempts in men than women as in all the age groups.

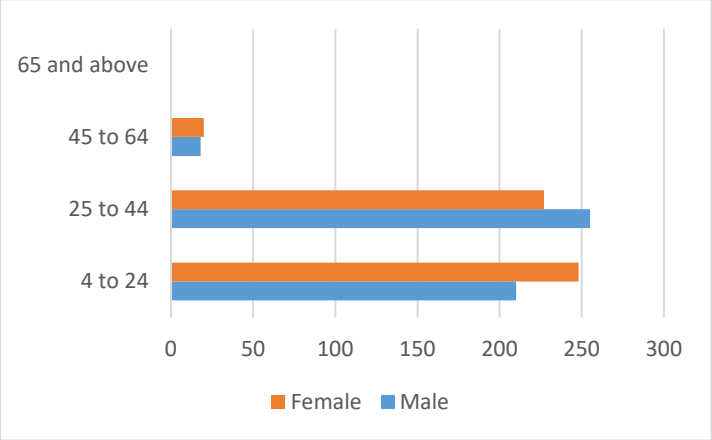


Figure 3. Gender and age representations of the participants in this study

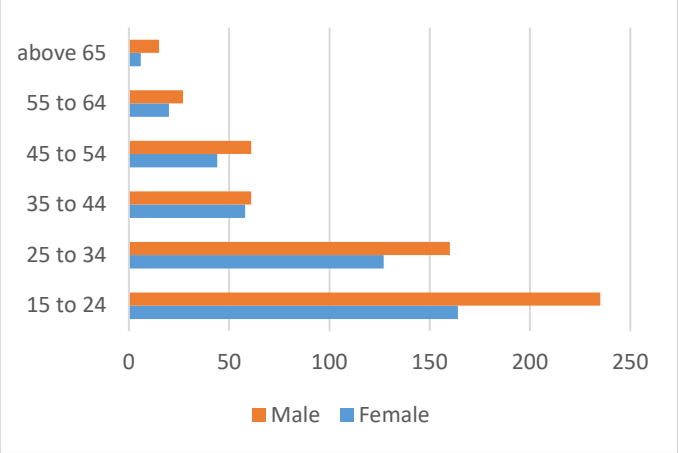


Figure 4. Age and gender representation of suicide or attempted victims by reported by the participants

Further, we collected data on the age range of the victims of suicide as reported by the participants. With this data, Figure 4 represents the age ranges of the suicide victims reported by the participants in this study. In the figure, it can be deduced that suicide is prevalent in the age of 15 to 24 years. In line with these findings, WHO in 2015 reported suicide to be prevalent among people in the age group of 15 to 29 [57]. Osafo *et al.* [58] found, in Ghana, that suicide behaviour is motivated by social taunting, hopelessness, and partner's infidelity. The American foundation for suicide prevention has, in real time, rated suicide as the 10th recorded deaths in the United State of America, while in Ghana suicide is the second highest deaths [57]. The difference in suicide deaths with respect to these countries could be attributed to the diverse characteristics of each context where culture, social, economic and political colouring accounts for the variations.

4.2. Emotional impact of suicide on witnesses

In this section, we present the findings arising from the inter-coders' agreements of emotions in the corpus. In addition, these researchers report the findings from the machine learning prediction. The inter-coder agreement *kappa* score was computed from all the three coders. As indicated in Table 1, a kappa score of 0.75 **(75%)** was obtained, which therefore indicates the variations in the coding patterns among the coders. The results from the coding is illustrated in Figure 5 and this is based on the predefined emotion categories *where 1- worry, 2- anger, 3 – fear, 4 – sadness, 5 – depression, 6 – surprise and 7- no emotion categories.* After obtaining the *kappa* score, the disagreements of the various instances of the data were subjected to further review and scrutiny. The deliberations among these researchers and the coders were meticulously carried out. In the end, we managed to agree on the instances of the data that were disagreed among the coders during the initial coding. The modified inter-coders' agreement of the emotions is shown in Table 1 while the initial coding by the coders is illustrated in Figure 5. The modified coded data was used as a *gold standard* for comparing with the machine learning prediction, as earlier established.

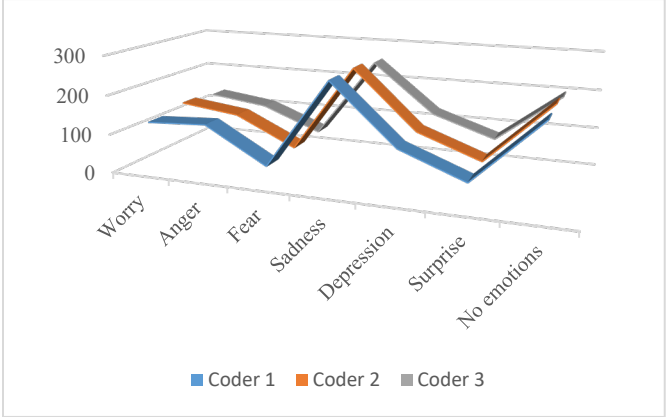


Figure 5. Coding patterns of the participants according to the emotion categories

Table 1: Inter-coded instance of emotions in the corpus

#	Emotion	# of coded instance
1	Worry	120 (12%)
2	Anger	138 (14%)
3	Fear	61 (6%)
4	**Sadness**	**251 (26%)**
5	Depression	131 (13%)
6	Surprise	70 (7%)
7	**No emotion**	**206 (21%)**
	TOTAL	**978 (100%)**

Young *et al.* [61] deduced that grief, sadness, and disbelief typical of all grief, overwhelming guilt, confusion, rejection, shame, and anger are also often prominent in suicide. From Table 1, the highest coded emotion is *sadness* (26% of the 978 instances*)*. In each case of the coding, *sadness* appeared as the dominant emotional impact on the participants (witnesses). This is followed by *no emotion* category. The implication is that the majority of the witnesses felt *sad* after they had witnessed suicide or suicide attempts. However, the *no emotion* category, which appeared as the second highest, implies that the witnesses did not feel any

emotional impact after witnessing suicide and suicide attempts. This could also mean that the participants declined to share their emotional antecedents as requested during the data collection. The *no emotion* indicates 21% of the entire text instances. This implies that of the 978 participants who accounted for witnessing suicide and suicide attempts, 206 of them did not express any emotions.

From Table 1, the lowest of the emotion categories is *fear*. Fear accounted for 6% in the participants' responses regarding the expression of emotions. The implication of this is that only 61 out of the 978 of the participants expressed *fear* after they had witnessed suicides or suicide attempts. The *fear* emotion category may have been aroused on the grounds that the participants assumed to have been those committing or attempting the suicide. This situation is likely to trigger fear, and could eventually lead to another form of emotion category, such as depression.

4.3. Classifiers performance

Having predicted the emotional impact of suicide from the witnesses' accounts, there was the need to ascertain the performance of the various classifiers used in this work. Table 2 shows the Precision, Recall and F-measure of the various classifiers as elaborated in the earlier section. From the table, the performances of the machine learning classifiers for each of the emotion categories are reported, including the overall performance of the classifiers. Aside the *no emotion* category, the highest predicted emotion by the SVM classifier is *depression*. To add is that, the SVM classifier yielded 90% accuracy for predicting depression. With this accuracy, the SVM produced 92% Recall and 87% Precision from predicting *depression* in the text corpus. However, *worry* was the lowest of the emotion categories which was predicted by the SVM classifier though this exceeded the general acceptable threshold of 70%. This implies that, of the 131 instances that expressed *depression*, 92% of them were identified and tagged, however only 87% of the identified proportion of the instances were accurately tagged and matched with the *gold standard*. This score is higher than the threshold of the inter-coder agreement of the 75% reported earlier.

In the same vein, the highest individual emotion category predicted by the MNB classifier is *depression*. From Table 2, the MNB classifier performed satisfactorily as the Recall, Precision and the F-measure yielded 88%, 90% and 89% respectively. These scores are beyond the generally acceptable 70% threshold. The implication is that while 88% of the 131 instances of the data was identified and recognised by the classifier, 90% was accurately predicted as the *gold standard* instance of the corpus. Just like the SVM, *worry* was found to be the lowest in terms of predicting emotions in the text.

In the nutshell, the best performing classifier is the SVM (accuracy = 78%), of which the Precision, Recall and F-measure exceeded the acceptable threshold of 70%. Although the difference in the performance is not significant, both classifiers can be concluded as suitable for predicting emotions in the witnesses' accounts with regards to suicide or suicide attempts.

Although the two classifiers had scores greater than the threshold, the SVM classifier was superior over the MNB. We conclude that machine learning, as compared with humans, can predict emotions in text that can be likened to humans.

Table 2: Evaluation results of the various machine learning predictions

	SVM			MNB		
Emotion	**R**	**P**	**F**	**R**	**P**	**F**
Worry	73	67	70	67	69	68
Anger	82	75	78	73	80	76
Fear	76	89	72	72	71	72
Sadness	78	72	75	68	65	67
Depression	**92**	**87**	**90**	**88**	**90**	**89**
Surprise	74	77	75	78	74	76
No emotion	**93**	**94**	**94**	89	90	90
TOTAL	**81**	**80**	**79**	**66**	**67**	**67**

5. Conclusion

We investigated the emotional impact of suicide on incidental witnesses using Machine Learning classification techniques. Usually, this domain of research is overlooked as many researchers focus on the cause and impact of suicide behaviour. To do this, we collected subjective data about some suicidal antecedents of people from a random setting. The collected data was pre-processed into corpus for suitability and easy annotation with predefined emotion categories adopted from Plutchik. The corpus, which was coded with the predefined emotion categories, was further analysed using Machine Learning classifiers – SVM and MNB. We employed 10-folds cross validation strategy in the classification of emotions in the corpus.

With regard to the manual analysis of the emotions in the corpus by well-trained coders, *sadness* emerged as the most salient emotion expressed by the participants (witnesses) while *fear* was found to be the lowest emotion category expressed by the participants (witnesses). However, the machine learning prediction yielded highest in predicting depression with insignificant variations in the other emotional categories. This nonetheless shows that people who have witnessed suicide or suicidal attempts are inherently affected by some form of emotions that may require urgent attention to alleviate. In addition, through the witnesses' accounts, the study revealed that suicide and suicidal attempts occur in the age of 15 to 24, with men being the highest victim.

By evaluating the performance of the classifiers, the SVM classifier emerged as superior in terms of predicting the emotional impact of suicide on the witnesses (participants). By comparing the performance of the various classifiers with humans, we conclude that the Machine Learning algorithms can predict emotions in text as humans. The outcome of the study contributes to the pool of research that sought to advocate for the use of machine learning for predicting social phenomenon, especially in text, for complementing decisions. Additionally, the finding has demonstrated the emotional impact of suicide on incidental witnesses.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] World Health Organization. World health statistics 2015. World Health Organization, 2015.
[2] UNICEF. Study of prevalence and dynamics of suicide among children and young people (12-24 years of age) in SUGHD REGION, TAJIKISTAN. 2013, pp. 1-88.

[3] R.M. Berman, A. Cappiello, A. Anand, D.A. Oren, G.R. Heninger, D.S. Charney & J.H. Krystal. "Antidepressant effects of ketamine in depressed patients." Biol Psychiatry. 47(4):351-4, 2000.

[4] W. Eisele, D.T. Reay & A. Cook. "Sites of Suicidal Gunshot Wounds", J Forensic Science, 26 (3), 1981.

[5] D. Klonsky & M.M Alexis. "The Three-Step Theory (3ST): A New Theory of Suicide Rooted in the Ideation-to-Action. Framework". International Journal of Cognitive Therapy, 8(2), 114–129, 2015.

[6] HKJC Centre for Suicide Research and Prevention, University of Hong Kong. Method Used in Completed Suicide, 2006. Available at http://csrp.hku.hk/WEB/eng/statistics.asp#3.

[7] J.J. Card. "Lethality of suicidal methods and suicide risk: Two distinct concepts", Omega 5, 1974.

[8] V. Admiraal,. "Guide to a Humane Self-Chosen Death Alt Suicide Holiday". Accessed on 24th April 2018 on http://ash2.wikkii.com/wiki/Helium, with their information taken from a range of sources but including, 2006.

[9] P. Nitschke, & F. Stewart. "The Peaceful Pill eHandbook", 2009.

[10] P. Passos & J.Y Chow, "Interpersonal coordination in team sports," in *Interpersonal Coordination and Performance in Social Systems*, ed. J. Y. Chow (New York, NY: Routledge), 154–164, 2016.

[11] M.J. Just, L. Pan, V. L. Cherkassky, D.L. McMakin,, C. Cha., M.K. Nock & D. Brent. "Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth". Nature Human Behavior, 2017. October 30. doi:10.1038/s41562-017-0234-y

[12] E.O. Wilson. "The Meaning of Human Existence". 1ST ed. W. W. Norton & Company, 2014, 192pages.

[13] G.A. Van-Kleef, A. Cheshin, A. H. Fischer & I.K Schneider.. "Editorial: The social nature of emotions". *Frontiers in Psychology*, *7*(896), 1–5, 2016.

[14] D. Keltner & J. Haidt. "Social functions of emotions at four levels of analysis". *Cognition and Emotion* 13: 505-21, 1999.

[15] P.M. Niedenthal & M. Brauer. "Social Functionality of Human Emotion" Annual Review of Psychology, Vol. 63, pp. 259-285, 2012.

[16] A.S. Morris, J.S. Silk, L. Steinberg, S.S. Myers & L.R. Robinson. "The Role of the Family Context in the Development of Emotion Regulation". Social Development. Author manuscript; Published in final edited form as: Soc Dev. 16(2): 361–38, 2007.

[17] D. Keltner & J.S Lerner. "Emotion." In The handbook of social psychology, ed. DT Gilbert, ST Fiske, G Lindzey, pp. 317-52. New York, NY: Wiley, 2010.

[18] P. Ekman. "Emotions revealed: Recognizing faces and feelings to improve communication and emotional life". New York, NY: Holt, 2007.

[19] E.A. Kolog, C.S. Montero & M. Tukiainen. "Development and Evaluation of automated e-counselling system for emotion and sentiment analysis". The electronic journal of information systems evaluation. 21(1), 1-19, 2018.

[20] D. Keltner, K. Oatley, J.M. Jenkins. "Understanding emotions". Hoboken, NJ: Wiley, 2004.

[21] R. Coughlan & T. Connolly. "Predicting affective responses to unexpected outcomes." Organizational Behavior and Human Decision Processes, 85(2), 211-225, 2001.

[22] B.A. Mellers, "Choice and the relative pleasure of consequences". *Psychological Bulletin* 126: 910-24, 2001.

[23] N.H. Frijda & B. Mesquita. "The social roles and functions of emotions". In S. Kitayama & H. R. Markus (Eds.), *Emotion and culture: Empirical studies of mutual influence.* Washington, DC: American Psychological Association, 1994.

[24] B. Parkinson. "Emotions are social". *Br. J. Psychol.* 1996, 87, 663–683. doi: 10.1111/j.2044-8295.1996.tb02615.x

[25] N.E. Dowd, D.G. Singer & R.F. Wilson, eds. "Handbook of Children, Culture, and Violence". California: Sage, 2006.

[26] D.A.Brent, J. Perper, G. Moritz, A. Friend, J. Schweers, C. Allman, L. McQuiston,M.B. Boylan, C. Roth, & L. Balach. "Adolescent witnesses to a peer suicide". Journal of American Academy of Child and Adolescent Psychiatry, *32*(6): 1184-1188, 1993.

[27] I.T. Young, A. Iglewicz, D. Glorioso, N. Lanouette, K. Seay, M. Ilapakurti, & S. Zisook. Suicide Bereavement and Complicated Grief. *Dialogues in Clinical Neuroscience*, *14*(2): 177–186, 2012.

[28] P. Goldblum, D. L. Espelage, J. Chu, & B. Bongar eds.. "Youth Suicide and Bullying: Challenges and Strategies for Prevention and Intervention". Oxford: Oxford University Press. 2015.

[29] F.B. Ribeiro, E.A.T. Lanna, M.A.D Bomfim, J.L. Donzele, M. Quadros, P. Cunha & S.L de. "True and apparent digestibility of protein and amino acids of feed in Nile tilapia". Rev. Bras. Zootec., 40 (5): 939-946, 2011

[30] K. Posner, G.K. Brown, B. Stanley, D.A. Brent, K.V. Yershova, M.A. Oquendo, G.W. Currier, G.A. Melvin, L. Greenhill, S. Shen & J.J. Mann. "The Columbia-Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults". Am J Psychiatry., 168(12):1266-77, 2011.

[31] M.A. Oquendo, G.M. Sullivan, S. Katherin , E. Baca-Garcia, B.H. Stanley, M.E. Sublette & J.J. Mann. "Toward a Biosignature for Suicide". The American Journal of Psychaitry , 171 (12,), pp. 1259-1277, 2014.

[32] E.A. Selby, M.A. Anestis, W. Theodore & E. J. Thomas E. J. "An Exploration of the Emotional Cascade Model in Borderline Personality Disorder" J Abnorm Psychol., 118(2): 375–38, 2009.

[33] N. Garnefski, V. Kraaij, & P. Spinhoven. Negative Life Events Cognitive Emotional Regulation and Emotional Problems. Personality and Individual Differences, 30, 1311-1327, 2001.

[34] R.L. Tamas, D. Menkes & R.S. El-Mallakh. S"timulating research: a prospective, randomized, double-blind, sham-controlled study of slow transcranial magnetic stimulation in depressed bipolar patients." Journal of Neuropsychiatry Clin Neurosci;19:198–199, 2007.

[35] T. Joiner. "Why people die by suicide". Cambridge, MA, US: Harvard University Press; 2005

[36] K.A. Van Orden, T.K. Witte, K.C. Cukrowicz, S.R. Braithwaite, E.A. Selby & T.E. Joiner Jr. "The interpersonal theory of suicide". Psychological review, 117(2), 575, 2010.

[37] C. Zlotnick, T.M. Shea, K. Rosen, E. Simpson, K. Mulrenin, A. Begin, *et al. "*An affect-management group for women with posttraumatic stress disorder and histories of childhood sexual abuse". Journal of Traumatic Stress, 10, pp. 425-436, 1997.

[38] T.J. Lynch, D.W. Bell & R. Sordella. "Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib". *Engl Journal of Med.; 350: 2129–2139, 2004.*

[39] J.O. Christensen & S. Knardahl. "Work and headache: a prospective study of psychological, social, and mechanical predictors of headache severity". Pain, 153(10):2119–2132, 2012.

[40] I. Lerner, B.C. Armstrong, & R. Frost. "What can we learn from learning models about sensitivity to letter-order in visual word recognition?" Journal of Memory and Language, 77, 40–58, 2014.

[41] M.D. Anestis, C.L. Bagge, M.T. Tull, & T.E. Joiner. "Clarifying the role of Emotion deregulation in the interpersonal-psychological theory of suicidal behaviour in an undergraduate sample". Journal of Psychiatric Research*, 45,* 603-611, 2011.

[42] S.P. Whiteside & D.R. Lynam. "The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity." Personality and Individual Differences*,, 30*(4), 669-689. 2001. doi: 10.1016/S0191-8869(00)00064-7

[43] M.D. Anestis & T.E. Joiner. "Examining the role of emotion in suicidality: Negative urgency as an amplifier of the relationship between components of the interpersonal–psychological theory of suicidal behavior and lifetime number of suicide attempts". Journal of affective disorders, *129*(1), 261-269, 2011.

[44] Baumeister, R. F. "Suicide as escape from self". Psychological Review*,* 97(1), 90-113, 1990.

[45] A.L. Thomas. "Suicidal Thoughts and Behaviour among Black College Students: Examining the Impact of Distress Tolerance and Social Support on Suicidality." Masters Theses & Specialist Projects, 2015.

[46] K. Rajappa, M. Gallagher, & R. Miranda. "Emotion dysregulation and vulnerability to suicidal ideation and attempts." Cognitive Therapy Res, 2012*,* 36-833-839. DOI:10.1007/s10608-011-9419-2

[47] B.L Fredrickson. "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions". American Psychologist*,, 56,* 218-226. 2011.

[48] E.A. Kolog, C. Suero Montero, & e. Sutinen. "Annotation agreement of emotions in text: The influence of counsellors emotional state on their emotion perception". In proceedings of international conference on advance learning technologies 2016.

[49] E.A. Kolog, & C. Suero Montero. "Towards automated e-counselling system based on counsellors' emotion perception". Education and information technologies, pp. 1-23, 2018.

[50] M. Hall, E. Frank., G. Holmes, B. Pfahringer, P. Reutemann, & I.H. Witten, "The WEKA Data Mining Software: An Update. SIGKDD Explorations", 2009.

[51] E.A. Kolog, C.S. Montero & T. Toivonen. "Using machine learning for sentiment and social influence analysis in text". In: Rocha, A., Guarda, T. (eds.) Proceedings of the International Conference on Information Technology & Systems (ICITS 2018). AISC, vol. 721, pp. 453–463. Springer, Cham (2018).

[52] A. Tripathy, A. Anand & S.K. Rath. "Document-level sentiment classification using hybrid machine learning approach". Knowledge and Information Systems. 2017; 53(3):805-31.

[53] Cohen, J. "Coefficient of agreement for nominal scales." Educational and Psychological *Measurement, 20*(1), 37–46, 1960.

[54] E. A. Kolog. Detecting Emotions in Students' Generated Content: An Evaluation of EmoTect System. In *International Conference on Technology in Education*, 2018, pp. 235-248. Springer, Singapore, 2018.

[55] N. Ibrahim, N. Amit, N.C. Din & H.C. Ong. Gender differences and psychological factors associated with suicidal ideation among youth in Malaysia. Psychology research and behaviour management, *10*, 129, 2017.
[56] American foundation for suicide prevention (AFSP). "Suicide statistics". Available on <https://afsp.org/about-suicide/suicide-statistics/> Accessed on the 13th April 2018.
[57] World Health Organization. "World health statistics 2016: monitoring health for the SDGs sustainable development goals". World Health Organization, 2016.
[58] J. Osafo, C.S. Akotia, J. Andoh-Arthur & E.N.B. Quarshie, "Attempted suicide in Ghana: Motivation, stigma, and coping." Death studies, 39(5), 274-280. 2015.
[59] R. Plutchik. Emotion: Theory, research, and experience: Theories of emotion. New York: Academic, 1, 399, 1980.
[60] L. Loyo, E. Martínez-Velázquez & J. Ramos-Loyo. "Influence of emotions on executive functions in suicide attempters." Suicidal Online, *4*(1), 42-55, 2013.